

VISIT...

LANZAROTE  
*Caliente*.COM

Solving Least Squares Problems  
Charles L. Lawson  
Richard J. Hanson  
Prentice-Hall, Inc.  
Englewood Cliffs, New Jersey

Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов/Пер. с англ. — М.: Наука. Гл. ред. физ.-мат. лит., 1986. — 232 с.

Книга посвящена изложению численных решений линейных задач метода наименьших квадратов. Достоинством книги являются: отбор наиболее устойчивых методов, полный анализ устойчивости, рассмотрение среднеквадратичных задач с линейными ограничениями, обзор методов перестройки ортогональных разложений при добавлении или удалении одного или нескольких наблюдений.

Для специалистов по прикладной математике, инженеров, а также для студентов и аспирантов.

Рецензент доктор физико-математических наук В.Н. Кубановская

© Prentice-Hall, Inc., Englewood Cliffs, N.J., 1974

© Издательство "Наука",  
Главная редакция  
физико-математической  
литературы,  
перевод на русский язык,  
предисловие и послесловие  
переводчика, 1986

Л 1702070000-070 15-86  
053 (02)-86

## ОГЛАВЛЕНИЕ

|                                                                                                  |    |
|--------------------------------------------------------------------------------------------------|----|
| ПРЕДИСЛОВИЕ ПЕРЕВОДЧИКА .....                                                                    | 5  |
| ПРЕДИСЛОВИЕ .....                                                                                | 6  |
| ГЛАВА 1. Введение .....                                                                          | 7  |
| ГЛАВА 2. Анализ задачи наименьших квадратов .....                                                | 10 |
| ГЛАВА 3. Ортогональное разложение посредством элементарных ортогональных преобразований .....    | 12 |
| ГЛАВА 4. Ортогональное разложение посредством сингулярного разложения .....                      | 18 |
| ГЛАВА 5. Теоремы о возмущениях сингулярных чисел .....                                           | 21 |
| ГЛАВА 6. Оценки для числа обусловленности треугольной матрицы .....                              | 24 |
| ГЛАВА 7. Псевдообратная матрица .....                                                            | 31 |
| ГЛАВА 8. Оценки возмущений для псевдообратных матриц .....                                       | 33 |
| ГЛАВА 9. Оценки возмущений для решений задачи НК .....                                           | 39 |
| ГЛАВА 10. Вычисления, использующие элементарные ортогональные преобразования .....               | 42 |
| ГЛАВА 11. Вычисление решения переопределенной или точно определенной задачи полного ранга .....  | 50 |
| ГЛАВА 12. Вычисление ковариационной матрицы решения .....                                        | 52 |
| ГЛАВА 13. Вычисление решения недоопределенной задачи полного ранга .....                         | 57 |
| ГЛАВА 14. Вычисление решения задачи НК, возможно, неполного псевдоранга .....                    | 59 |
| ГЛАВА 15. Анализ погрешностей округлений для преобразований Хаусхолдера .....                    | 63 |
| ГЛАВА 16. Анализ погрешностей округлений для задачи НК .....                                     | 69 |
| ГЛАВА 17. Анализ погрешностей округлений для задачи НК в арифметике со смешанной точностью ..... | 76 |
| ГЛАВА 18. Вычисление сингулярного разложения и решение задачи НК .....                           | 81 |
| § 1. Введение .....                                                                              | 81 |
| § 2. QR-алгоритм для симметричных матриц .....                                                   | 82 |
| § 3. Вычисление сингулярного разложения .....                                                    | 83 |

|                                                                                                                                                                                 |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| § 4. Решение задачи НК посредством сингулярного разложения                                                                                                                      | 90  |
| § 5. Организация программы, вычисляющей сингулярное разложение                                                                                                                  | 91  |
| ГЛАВА 19. Другие методы для задачи наименьших квадратов                                                                                                                         | 92  |
| § 1. Нормальные уравнения и разложение Холесского                                                                                                                               | 93  |
| § 2. Модифицированная ортогонализация Грама-Шмидта                                                                                                                              | 99  |
| ГЛАВА 20. Линейные задачи наименьших квадратов с линейными ограничениями-равенствами: решение с помощью базиса нуль-пространства                                                | 103 |
| ГЛАВА 21. Линейные задачи наименьших квадратов с линейными ограничениями-равенствами: решение посредством прямого исключения                                                    | 111 |
| ГЛАВА 22. Линейные задачи наименьших квадратов с линейными ограничениями-равенствами: решение путем взвешивания                                                                 | 114 |
| ГЛАВА 23. Линейные задачи наименьших квадратов с линейными ограничениями-неравенствами                                                                                          | 122 |
| § 1. Введение                                                                                                                                                                   | 122 |
| § 2. Характеризация решения                                                                                                                                                     | 123 |
| § 3. Задача NNLS                                                                                                                                                                | 124 |
| § 4. Задача LDP                                                                                                                                                                 | 127 |
| § 5. Преобразование задачи НКН в задачу LDP                                                                                                                                     | 129 |
| § 6. Задача НКН с ограничениями-уравнениями                                                                                                                                     | 130 |
| § 7. Пример выравнивания при наличии ограничений                                                                                                                                | 131 |
| ГЛАВА 24. Модификация $QR$ -разложения матрицы при добавлении или удалении столбцов                                                                                             | 134 |
| ГЛАВА 25. Практический анализ задач метода наименьших квадратов                                                                                                                 | 137 |
| § 1. Общие соображения                                                                                                                                                          | 137 |
| § 2. Левое умножение $A$ и $b$ на матрицу $G$                                                                                                                                   | 140 |
| § 3. Правое умножение $A$ на матрицу $H$ и замена переменных $x = H\tilde{x} + \xi$                                                                                             | 141 |
| § 4. Приписывание дополнительных строк к $[A : b]$                                                                                                                              | 144 |
| § 5. Удаление переменных                                                                                                                                                        | 149 |
| § 6. Сингулярный анализ                                                                                                                                                         | 151 |
| ГЛАВА 26. Примеры некоторых методов анализа задачи наименьших квадратов                                                                                                         | 153 |
| ГЛАВА 27. Модификация $QR$ -разложения при добавлении или удалении строки (с приложениями к последовательной обработке задач с большими или ленточными матрицами коэффициентов) | 160 |
| § 1. Последовательное накапливание                                                                                                                                              | 161 |
| § 2. Последовательное накапливание ленточных матриц                                                                                                                             | 164 |
| § 3. Пример: линейные сплайны                                                                                                                                                   | 169 |
| § 4. Сглаживание посредством кубических сплайнов                                                                                                                                | 172 |
| § 5. Удаление строк                                                                                                                                                             | 174 |
| ПРИЛОЖЕНИЕ А. ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ                                                                                                                                           | 180 |
| ПРИЛОЖЕНИЕ В. ДОКАЗАТЕЛЬСТВО ГЛОБАЛЬНОЙ КВАДРАТИЧНОЙ СХОДИМОСТИ $QR$ -АЛГОРИТМА                                                                                                 | 186 |
| ПОСЛЕСЛОВИЕ. Х.Д. ИКРАМОВ                                                                                                                                                       | 194 |
| СПИСОК ЛИТЕРАТУРЫ                                                                                                                                                               | 219 |
| СПИСОК ЛИТЕРАТУРЫ, ДОБАВЛЕННЫЙ ПРИ ПЕРЕВОДЕ                                                                                                                                     | 226 |

## ПРЕДИСЛОВИЕ ПЕРЕВОДЧИКА

Опубликованная в 1974 г. книга Ч. Лоусона и Р. Хенсона имела двойное назначение. С одной стороны, она задумана как учебник по прямым методам решения линейных среднеквадратичных задач — для лиц, не знакомых с основами вычислительной алгебры, и как справочник — для тех, кто имеет опыт работы в этой области. С другой стороны, в книге содержались тексты фортранных подпрограмм описанных в ней методов.

Через 12 лет после выхода книги можно констатировать, что она полностью сохранила свое значение учебника-справочника, все еще единственного по данному предмету. В советской монографической литературе методы численного решения задач наименьших квадратов обходились вниманием. Исключением являются лишь отдельные главы книг В.В. Воеводина, например "Вычислительные основы линейной алгебры" (М.: Наука, 1977). Можно надеяться поэтому, что перевод данной книги, дополненный небольшим послесловием — комментарием к литературе последнего десятилетия, будет полезен советским читателям. В то же время было решено отказаться от публикации текстов программ по следующим соображениям. Для наиболее содержательного и сложного метода — алгоритма сингулярного разложения — хорошая программа на фортране имеется в книге [12\*]. Программы прочих методов легко получить компилицией обычных алгебраических процедур: ортогонально-треугольного разложения матрицы, решения треугольной системы и т.п. Соответствующие стандартные подпрограммы можно найти в любой сколько-нибудь развитой библиотеке прикладного математического обеспечения.

Книга в основном посвящена задачам с заполненными матрицами, размеры  $m$ ,  $n$  которых удовлетворяют условию  $mn < M$ , где  $M$  — оперативная память используемой ЭВМ. Для многих современных машин  $M$  составляет величину порядка нескольких мегаслов, и методы, изложенные в книге, позволяют решать достаточно большие задачи. Но для сверхбольших задач с сильной разреженностью условие  $mn < M$  оказывается обременительным, а хранение матрицы полным массивом — излишним. Из разреженных задач в книге рассматривается лишь специальная ситуация ленточных матриц. Более полная картина разреженного случая дана в обзоре [3\*].

При переводе сохранена непривычная для советского читателя сквозная нумерация формул, утверждений и определений.

Я хотел бы поблагодарить Д.С. Шмерлинг и В.Ф. Матвеева за помощь в переводе терминов регрессионного анализа.

Х.Д. Икрамов

## ПРЕДИСЛОВИЕ

В этой книге собрана информация о задачах метода наименьших квадратов и практических алгоритмах их решения, разработанных главным образом в течение последнего десятилетия. Эта информация будет полезна научно-му работнику, инженеру или студенту, связанным в своей работе с анализом и решением систем линейных алгебраических уравнений. Такая система может быть определена, переопределена или недоопределена; она может быть совместна или несовместна. Она может сопровождаться ограничениями в форме линейных уравнений или неравенств.

Специалисты конкретных областей разработали методы и терминологию для задач наименьших квадратов из своих дисциплин. Материал, представленный в нашей книге, может помочь в преодолении этой разобщенности и достижении методологического и терминологического единства.

По существу, все реальные задачи нелинейны. Многие методы анализа нелинейных задач или вычислений на основе нелинейных моделей включают в себя процедуру локальной замены нелинейной задачи линейной. В частности, различные методы анализа и решения нелинейных задач метода наименьших квадратов предполагают решение последовательности линейных задач. Важное требование этих методов — умение вычислять такие решения линейных задач наименьших квадратов (возможно, плохо обусловленных), которые были бы приемлемы в контексте нелинейной задачи.

Читателю, интересующемуся в первую очередь практическими приложениями, мы рекомендуем начать чтение книги с гл. 25.

Начиная с Гаусса, многие математики участвовали в теоретической и практической разработке техники решения задач метода наименьших квадратов. Мы хотим особенно выделить вклад профессора Дж. Голуба, которому в этой области принадлежит много важных идей и алгоритмов.

Первый из авторов познакомил второго с задачей наименьших квадратов в 1966 г. С тех пор авторы тесно сотрудничали в JPL<sup>\*)</sup>, адаптируя устойчивые математические методы к практическим вычислительным задачам. Данная книга является частью этой совместной работы.

Ч.Л. Лоусон  
Р.Дж. Хенсон

<sup>\*)</sup>JPL — Jet Propulsion Laboratory — Лаборатория реактивного движения. (Примеч. пер.)

## ГЛАВА I

### ВВЕДЕНИЕ

Эта книга задумана одновременно как учебник и как справочник для лиц, которым приходится решать линейные задачи теории наименьших квадратов. Такие задачи часто возникают как составная часть некоторой более обширной вычислительной проблемы. Например, определение орбиты космического корабля нередко сводится математиками к решению многоточечной краевой задачи для обыкновенного дифференциального уравнения. При этом вычисление орбитальных параметров обычно требует нелинейного оценивания в смысле наименьших квадратов; в последнем используют различные схемы линеаризации.

Более общо, почти любая задача, в которой исходных данных достаточно для того, чтобы переопределить решение, требует применения того или иного метода аппроксимаций. Наиболее часто в качестве критерия аппроксимации выбирают метод наименьших квадратов.

При расчетах по методу наименьших квадратов часто встречаются дополнительные ограничения. Их стоит определить более точно, чтобы облегчить последующее построение алгоритмов. Опишем несколько примеров.

Задача может быть связана с некоторыми соотношениями равенства или неравенства между переменными. Она может требовать обработки столь большого объема информации, что главной проблемой становится распределение машинной памяти.

Во многих случаях цель вычислений по методу наименьших квадратов не исчерпывается тем, чтобы найти некоторый набор чисел, который "решает" задачу. Скорее, исследователь желает получить дополнительную количественную информацию, описывающую связь решения с исходными данными. В частности, задача может допускать целое семейство различных решений, которые почти в равной степени удовлетворяют поставленным условиям. Исследователь может пожелать описать эту неопределенность и затем произвести выбор в указанном семействе в соответствии с некоторыми дополнительными требованиями.

В этой книге представлены численные методы решения задач теории наименьших квадратов, учитывающие высказанные выше положения. Эти методы успешно использовались большим коллективом инженеров и научных сотрудников, включающим авторов, в ходе выполнения программы NASA непилотируемых космических полетов.

Задача наименьших квадратов, которую мы здесь рассматриваем, в разных научных дисциплинах называется по-разному. Например, математики могут подойти к ней как к задаче отыскания для заданной точки



функционального пространства ближайшей точки в заданном подпространстве. Специалисты по численному анализу тоже нередко использовали этот подход, при котором в тени остается вопрос об ошибках входной информации. Вследствие этого теряется возможность извлечь выгоду из произвола, часто присутствующего в решении.

Статистики вводят в свою постановку задачи вероятностные распределения и используют для описания этой области термины типа *регрессионный анализ*. Инженеры приходят к этой задаче, занимаясь такими предметами, как *оценивание параметров* или *фильтрация*.

Главное состоит в следующем: когда эти задачи (сформулированные в любом из названных контекстов) достигают стадии конкретных расчетов, они содержат в себе одну и ту же центральную проблему, а именно последовательность линейных задач наименьших квадратов.

Эту основную линейную задачу наименьших квадратов (НК) можно сформулировать следующим образом:

**Задача НК.** Пусть даны действительная  $m \times n$ -матрица  $A$  ранга  $k \leq \min(m, n)$  и действительный  $m$ -вектор  $b$ . Найти действительный  $n$ -вектор  $x_0$ , минимизирующий евклидову длину вектора  $Ax - b$ .

(В приложении А читатель может найти определения незнакомых ему терминов линейной алгебры.) Для обозначения задачи НК мы будем использовать символику  $Ax \approx b$ .

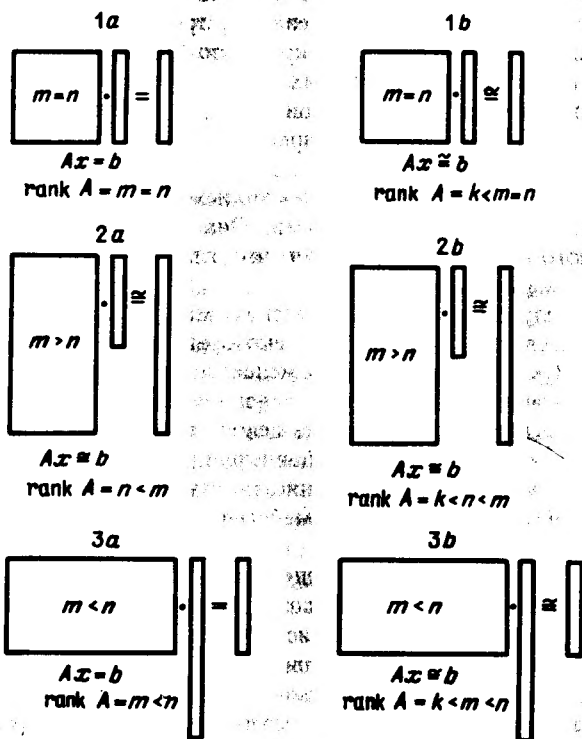


Рис. 1.1. Шесть случаев задачи НК в соответствии со сравнительной величиной  $m, n$  и ранга  $A$ .

Эту задачу можно поставить и исследовать также и для комплексных  $A$  и  $b$ . Комплексный случай встречается на практике гораздо реже, чем действительный. В то же время теория и численные методы для действительного случая непосредственно переносятся на комплексный.

Помимо приведенной формулировки задачи НК (в данном контексте) есть еще дополнительное условие: числовые данные, составляющие  $A$  и  $b$ , имеют лишь конечное число верных разрядов; последующие разряды совершенно не определены и, следовательно, произвольны. Это обычное для практики положение дел. Оно связано, в частности, с ограниченной точностью измерений или наблюдений. Важно в должной мере учесть эту ситуацию и с выгодой использовать ее для получения подходящего приближенного решения задачи. Мы будем обсуждать методы, позволяющие достигнуть этого, в особенности в гл. 25, 26.

Рассмотрим в качестве важного примера случай, когда линейная задача наименьших квадратов происходит из нелинейной и вектор решения  $x_0$  должен использоваться как поправка, прибавляемая к текущему номинальному решению нелинейной задачи. Линеаризованная задача будет полезным приближением к нелинейной лишь в некоторой ограниченной окрестности. Если имеются различные векторы, дающие достаточно малые невязки в линейной задаче, то можно предпочесть тот, что имеет наименьшую длину. Это увеличивает вероятность остаться в окрестности, где разумно линейное приближение.

Следующие три положения лежат в основе нашего подхода к методу наименьших квадратов:

1. Поскольку исходные данные задачи НК не вполне определены, мы можем изменить их, приспособив к своим нуждам.
2. Мы будем применять ортогональные матрицы исключения непосредственно к линейной системе задачи НК.
3. Всюду большое внимание уделяется практичности при машинной реализации.

Дадим краткий комментарий к первым двум положениям. Наша цель при изменении задачи, о котором говорится в п. 1, состоит в том, чтобы избежать ситуации (см. выше), когда "малое" изменение исходных данных приводит к "большим" изменениям решения.

Матрицы ортогональных преобразований, упомянутые в п. 2, находят естественное место в методе наименьших квадратов, поскольку они оставляют неизменной евклидову длину вектора. Кроме того, их использование желательно и вследствие присущей им численной устойчивости по отношению к распространению ошибок или неопределенности входных данных.

Мы не выдвигали никаких предположений относительно сравнительной величины параметров  $m$  и  $n$ . Для последующего обсуждения задачи НК удобно разделить шесть случаев, иллюстрируемых на рис. 1.1.

Наибольшее внимание в этой книге будет направлено на случай 2a; при этом специально будет рассмотрена ситуация, когда неопределенность во входных данных приводит к случаю 2b. Однако алгоритмы и их обсуждение будут даны для всех шести случаев.

# АНАЛИЗ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ

Основным содержанием этой главы будет анализ задачи НК, основанный на представлении  $m \times n$ -матрицы  $A$  произведением вида  $HRK^T$ , где  $H$  и  $K$  — ортогональные матрицы. Определения ортогональной матрицы и других понятий линейной алгебры приведены в приложении А.

Наш интерес к разложению  $A = HRK^T$  общего вида мотивирован практичностью и полезностью некоторых конкретных вычислимых разложений этого типа, которые будут введены в гл. 3, 4.

Важным свойством ортогональных матриц является сохранение евклидовой длины при умножении. Это значит, что для любого  $m$ -вектора  $u$  и любой ортогональной  $m \times m$ -матрицы  $Q$

$$\|Qu\| = \|u\|. \quad (2.1)$$

В контексте задачи НК, где речь идет о минимизации евклидовой длины вектора  $Ax - b$ , имеем

$$\|Q(Ax - b)\| = \|QA x - Qb\| = \|Ax - b\| \quad (2.2)$$

для произвольной ортогональной  $m \times m$ -матрицы  $Q$  и любого  $n$ -вектора  $x$ .

Использование ортогональных преобразований позволяет выразить решение задачи НК следующим образом:

**Теорема 2.3.** Пусть  $A$  —  $m \times n$ -матрица ранга  $k$ , представленная в виде  $A = HRK^T$ ,

где  $H$  — ортогональная  $m \times m$ -матрица;  $R$  —  $m \times n$ -матрица вида  $R = \begin{bmatrix} R_{11} & 0 \\ 0 & 0 \end{bmatrix}$ ;  $R_{11}$  —  $k \times k$ -матрица ранга  $k$ ;  $K$  — ортогональная  $n \times n$ -матрица. Определим вектор

$$H^T b = g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix} \quad (2.5)$$

и введем новую переменную

$$K^T x = y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \quad (2.6)$$

Определим  $\tilde{y}_1$  как единственное решение системы

$$R_{11} y_1 = g_1. \quad (2.7)$$

Тогда:

1) Все решения задачи о минимизации  $\|Ax - b\|$  имеют вид

$$\hat{x} = K \begin{bmatrix} \tilde{y}_1 \\ y_2 \end{bmatrix}, \quad (2.8)$$

где  $y_2$  произвольно.

2) Любой такой вектор  $\hat{x}$  приводит к одному и тому же вектору невязки  $r$ :

$$r = b - A\hat{x} = H \begin{bmatrix} 0 \\ g_2 \end{bmatrix}. \quad (2.9)$$

3) Для нормы  $r$  справедливо

$$\|r\| = \|b - A\hat{x}\| = \|g_2\|. \quad (2.10)$$

4) Единственным решением минимальной длины является вектор

$$\tilde{x} = K \begin{bmatrix} \tilde{y}_1 \\ 0 \end{bmatrix}. \quad (2.11)$$

**Доказательство.** Заменяя  $A$  правой частью равенства (2.4) и применяя (2.2), получаем

$$\|Ax - b\|^2 = \|HRK^T x - b\|^2 = \|RK^T x - H^T b\|^2. \quad (2.12)$$

Из уравнений (2.5) — (2.8) следует, что

$$\|Ax - b\|^2 = \|R_{11} y_1 - g_1\|^2 + \|g_2\|^2 \quad (2.13)$$

для всех  $x$ .

Правая часть (2.13) имеет минимальное значение  $\|g_2\|^2$ , если

$$R_{11} y_1 = g_1. \quad (2.14)$$

Уравнение (2.14) допускает единственное решение  $\tilde{y}_1$ , так как ранг  $R_{11}$  равен  $k$ .

Общее решение  $y$  выражается формулой

$$\tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ y_2 \end{bmatrix}, \quad (2.15)$$

где  $y_2$  произвольно.

Для вектора  $\hat{x}$ , определенного уравнением (2.8), имеем

$$b - A\hat{x} = b - HRK^T \hat{x} = H(g - R\tilde{y}) = H \begin{bmatrix} 0 \\ g_2 \end{bmatrix},$$

что устанавливает равенство (2.9).

Ясно, что среди векторов  $\tilde{y}$  вида (2.15) наименьшую длину имеет тот, для которого  $y_2 = 0$ . Из (2.8) следует, что решением наименьшей евклидовой длины будет вектор

$$\tilde{x} = K \begin{bmatrix} \tilde{y}_1 \\ 0 \end{bmatrix}. \quad (2.16)$$

Теорема 2.3 доказана.

Всякое разложение  $m \times n$ -матрицы  $A$  того же типа, что и (2.4), мы будем называть ортогональным разложением  $A$ .

В случае  $k = n$  или  $k = m$  величины с размерностями соответственно  $n - k$  или  $m - k$  отсутствуют. В частности, при  $k = n$  решение задачи НК единственно.

Заметим, что решение минимальной длины, множество всех решений и минимальное значение для задачи минимизации  $\|Ax - b\|$  определяются единственным образом. Они не зависят от конкретного ортогонального разложения.

Упражнение

2.17. Пусть  $A - m \times n$ -матрица ( $m > n$ ) ранга  $k$  и  $Q_1 R_1 = A = Q_2 R_2$ . Если  $Q_i - m \times n$ -матрицы с ортогональными столбцами, а  $R_i -$  верхние треугольные  $n \times n$ -матрицы, то существует диагональная матрица  $D$  с диагональными элементами, равными  $+1$  или  $-1$ , такая, что  $Q_2 D = Q_1$  и  $DR_1 = R_2$ .

### ГЛАВА 3

#### ОРТОГОНАЛЬНОЕ РАЗЛОЖЕНИЕ ПОСРЕДСТВОМ ЭЛЕМЕНТАРНЫХ ОРТОГОНАЛЬНЫХ ПРЕОБРАЗОВАНИЙ

В гл. 2 показано, что задача НК может быть решена, если имеется ортогональное разложение  $m \times n$ -матрицы  $A$ . В этой главе мы установим существование такого разложения для произвольной матрицы, используя явные ортогональные преобразования.

Два типа элементарных ортогональных преобразований, которые нам понадобятся, представлены в следующих двух леммах.

Л е м м а 3.1. Пусть дан (ненулевой)  $m$ -вектор  $v$ . Существует ортогональная матрица  $Q$  такая, что

$$Qv = -\sigma \|v\| e_1, \quad (3.2)$$

где

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \sigma = \begin{cases} +1, & \text{если } v_1 \geq 0, \\ -1, & \text{если } v_1 < 0, \end{cases}$$

а  $v_1$  — первая компонента вектора  $v$ .

Доказательство. Положим

$$u = v + \sigma \|v\| e_1, \quad Q = I_m - 2 \frac{uu^T}{u^T u}. \quad (3.3)$$

Доказательство завершается прямой проверкой симметричности и ортогональности матрицы  $Q$  из (3.3) и остающейся части утверждения леммы 3.1.

Преобразование, определяемое формулой (3.3), использовалось Хаусхолдером [101] для решения некоторых задач на собственные значения. По этой причине матрица  $Q$  и соответствующее преобразование называются матрицей и преобразованием Хаусхолдера.

Это преобразование с геометрической точки зрения есть отражение в  $m - 1$ -мерном подпространстве  $S$ , ортогональном к вектору  $u$ . Это значит, что  $Qu = -u$  и  $Qs = s$  для всех  $s \in S$ .

В специальном случае, когда нужно преобразовать к нулю только один элемент вектора  $v$ , часто используют следующее преобразование Ги-

венса [69]. Так как это преобразование изменяет только две компоненты вектора, то достаточно изучить его действие на 2-вектор.

Л е м м а 3.4. Пусть дан 2-вектор  $v = (v_1, v_2)^T$ , причем  $v_1 \neq 0$  либо  $v_2 \neq 0$ . Существует ортогональная  $2 \times 2$ -матрица

$$G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \quad (3.5)$$

такая, что

$$c^2 + s^2 = 1, \quad (3.6)$$

$$Gv = \begin{bmatrix} (v_1^2 + v_2^2)^{1/2} \\ 0 \end{bmatrix}. \quad (3.7)$$

Доказательство. Положим просто

$$c = \frac{v_1}{(v_1^2 + v_2^2)^{1/2}}, \quad (3.8)$$

$$s = \frac{v_2}{(v_1^2 + v_2^2)^{1/2}}. \quad (3.9)$$

Легко показать, что матрица  $G$  из (3.5) ортогональна и удовлетворяет условию (3.7). Лемма 3.4 доказана.

Заметим еще, что  $G$  можно выбрать не только ортогональной, но и симметричной, полагая

$$G = \begin{bmatrix} c & s \\ s & -c \end{bmatrix}; \quad (3.10)$$

с и  $s$  те же, что и в формулах (3.8), (3.9).

Т е о р е м а 3.11. Пусть  $A - m \times n$ -матрица. Существует ортогональная  $m \times m$ -матрица  $Q$  такая, что в матрице  $QA = R$  под главной диагональю стоят только нулевые элементы.

Выберем ортогональную  $m \times m$ -матрицу  $Q$  в соответствии с леммой 3.1 так, чтобы первый столбец  $Q_1 A$  имел нулевые компоненты со 2-й по  $m$ -ю. Далее выбираем ортогональную  $(m - 1) \times (m - 1)$ -матрицу  $P_2$  следующим образом. Будучи применена к  $m - 1$  вектору, составленному из компонент со 2-й по  $m$ -ю второго столбца матрицы  $Q_1 A$ , она аннулирует компоненты с 3-й по  $m$ -ю этого вектора. Матрица преобразования

$$Q_2 = \begin{bmatrix} 1 & 0 \\ 0 & P_2 \end{bmatrix}$$

ортогональна, и  $Q_2 Q_1 A$  имеет в первых двух столбцах нули под главной диагональю.

Продолжая таким образом, можно построить произведение, состоящее самое большее из  $n$  ортогональных преобразований, которое трансформирует  $A$  к верхней треугольной форме. Эти рассуждения можно формализовать и получить доказательство теоремы 3.11 методом конечной индукции.

Алгоритмические детали построения указанных преобразований будут даны в гл. 10. Полученное представление матрицы произведением ортогональной и верхней треугольной матриц называется  $QR$ -разложением. Оно

ры [59, 72].

Для случаев 1а и 2а рис. 1.1, где  $\text{rank } A = n$ , теорема (3.11) устанавливает существование ортогонального разложения  $A$ . Действительно, согласно теореме 3.11, можно написать

$$A = Q^T R = Q^T R I_n; \quad (3.12)$$

матрицы  $Q^T$ ,  $R$ ,  $I_n$  этого представления имеют свойства, требуемые в соответствии с теоремой (2.3) от сомножителей  $H$ ,  $R$ ,  $K^T$  ортогонального разложения матрицы  $A$ .

Если  $\text{rank } A = m$  (см. 3а на рис. 1.1), то теорема 3.11 позволяет нам написать

$$A^T = Q^T R, \quad (3.13)$$

так что

$$A = R^T Q = I_m R^T Q. \quad (3.14)$$

В этом представлении  $I_m$ ,  $R^T$ ,  $Q$  имеют свойства, требуемые в соответствии с теоремой 2.3 от сомножителей  $H$ ,  $R$ ,  $K^T$  ортогонального разложения матрицы ранга  $m$ .

Для случаев 1б, 2б и 3б рис. 1.1 матрица  $R$ , полученная в теореме 3.11, необязательно имеет форму, требуемую ортогональным разложением.

Мы переходим к обсуждению дополнительных преобразований, которые позволяют получить ортогональное разложение и для этих случаев.

**Теорема 3.15.** Пусть  $A$  —  $m \times n$ -матрица ранга  $k$ , причем  $k < n \leq m$ . Существуют ортогональная  $m \times m$ -матрица  $Q$  и  $n \times n$ -матрица перестановки  $P$  такие, что

$$QAP = \begin{bmatrix} R & T \\ 0 & 0 \end{bmatrix}, \quad (3.16)$$

где  $R$  — верхняя треугольная  $k \times k$ -матрица ранга  $k$ .

**Доказательство.** Выберем матрицу перестановки  $P$  таким образом, чтобы первые  $k$  столбцов матрицы  $AP$  были линейно независимы. Согласно теореме 3.11, найдется ортогональная  $m \times m$ -матрица  $Q$  такая, что  $QAP$  — верхняя треугольная. Поскольку первые  $k$  столбцов  $AP$  линейно независимы, это же будет верно для первых  $k$  столбцов  $QAP$ .

Все элементы матрицы  $QAP$ , стоящие на пересечении строк с номерами  $k+1, \dots, m$  и столбцов с номерами  $k+1, \dots, n$ , будут нулями. В противном случае  $\text{rank } QAP > k$ , что противоречит предположению  $\text{rank } A = k$ . Итак,  $QAP$  имеет форму, указанную правой частью (3.16). Теорема 3.15 доказана.

Подматрицу  $[R : T]$  из правой части (3.16) можно теперь преобразовать к компактной форме, требуемой от матрицы  $R$  из теоремы 2.3. Это преобразование описывает следующая лемма.

**Лемма 3.17.** Пусть  $[R : T]$  —  $k \times n$ -матрица, причем  $R$  имеет ранг  $k$ . Существует ортогональная  $n \times n$ -матрица  $W$  такая, что

$$[R : T]W = [\hat{R} : 0], \quad (3.18)$$

где  $\hat{R}$  — нижняя треугольная матрица ранга  $k$ .

Лемма 3.17 вытекает из теоремы 3.15, если отождествить величины  $n$ ,  $k$ ,  $[R : T]$ ,  $W$  из формулировки леммы с соответствующими величинами  $m$ ,  $n$ ,  $A^T$ ,  $Q^T$  теоремы 3.15.

Используя лемму 3.17 вместе с теоремой 3.15, можно доказать следующую теорему.

**Теорема 3.19.** Пусть  $A$  —  $m \times n$ -матрица ранга  $k$ . Найдутся ортогональная  $m \times m$ -матрица  $H$  и ортогональная  $n \times n$ -матрица  $K$  такие, что

$$H^T A K = R, \quad A = H R K^T, \quad (3.20)$$

где

$$R = \begin{bmatrix} R_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad (3.21)$$

причем  $R_{11}$  — невырожденная треугольная  $k \times k$ -матрица.

Заметим, что выбором  $H$  и  $K$  в уравнении (3.20) можно добиться, чтобы  $R_{11}$  в (3.21) была верхней или нижней треугольной.

Подведем итоги. Было показано, что существуют конструктивные процедуры для получения ортогональных разложений вида  $A = H R K^T$  для всех шести случаев, изображенных на рис. 1.1. Во всех случаях подматрица  $R_{11}$  ранга  $k$  из теоремы 2.3 будет получена в треугольной форме. Поэтому вычислить решение уравнения (2.7) будет совсем просто.

Как уже отмечалось, в случаях 1а и 2а в качестве матрицы  $K^T$  разложения можно взять единичную матрицу  $I_n$ . Точно так же в случае 3а в качестве  $H$  можно взять единичную матрицу  $I_m$ .

С целью иллюстрации этих ортогональных разложений мы приведем численный пример для каждого из шести случаев рис. 1.1.

**Случай 1а. Квадратная невырожденная матрица:**

$$m = n = 3, \quad \text{rank } A = 3, \quad QA = R. \quad (3.22)$$

$$Q = \begin{bmatrix} -0,4800 & -0,4129 & -0,7740 \\ 0,6616 & -0,7498 & -0,0103 \\ -0,5761 & -0,5170 & 0,6331 \end{bmatrix},$$

$$A = \begin{bmatrix} 0,4087 & 0,1593 & 0,6593 \\ 0,3515 & 0,9665 & 0,6245 \\ 0,6590 & 0,9343 & 0,9039 \end{bmatrix},$$

$$R = \begin{bmatrix} -0,8514 & -1,1987 & -1,2740 \\ 0,0 & -0,6289 & -0,0414 \\ 0,0 & 0,0 & -0,1304 \end{bmatrix}.$$

**Случай 2а. Матрица переопределена и имеет полный ранг:**

$$m = 3, \quad n = 2, \quad \text{rank } A = 2. \quad (3.23)$$

$$QA = \begin{bmatrix} R \\ 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} -0,4744 & -0,4993 & -0,7250 \\ 0,5840 & -0,7947 & 0,1652 \\ -0,6587 & -0,3450 & 0,6687 \end{bmatrix},$$

$$A = \begin{bmatrix} 0,4087 & 0,1594 \\ 0,4302 & 0,3516 \\ 0,6246 & 0,3384 \end{bmatrix},$$

$$R = \begin{bmatrix} -0,8615 & -0,4965 \\ 0,0 & -0,1304 \end{bmatrix}.$$

Случай 3а. Матрица недоопределена и имеет полный ранг:  
 $m = 2, n = 3, \text{rank } A = 2.$  (3.24)

$$AQ = [R : 0],$$

$$A = \begin{bmatrix} 0,4087 & 0,4301 & 0,6246 \\ 0,1594 & 0,3515 & 0,3384 \end{bmatrix},$$

$$Q = \begin{bmatrix} -0,4744 & 0,5840 & -0,6587 \\ -0,4993 & -0,7947 & -0,3450 \\ -0,7250 & 0,1652 & 0,6687 \end{bmatrix},$$

$$R = \begin{bmatrix} -0,8615 & 0,0 \\ -0,4965 & -0,1304 \end{bmatrix}.$$

Случай 1б. Квадратная вырожденная матрица:

$$m = n = 5, \text{rank } A = 3.$$

$$QAK = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix},$$

(3.25)

$$Q = \begin{bmatrix} -0,0986 & -0,6926 & -0,5937 & -0,0522 & -0,3941 \\ 0,3748 & 0,0072 & -0,5424 & 0,3529 & 0,6639 \\ -0,1244 & 0,2438 & -0,3484 & -0,8633 & 0,2419 \\ 0,2401 & -0,6653 & 0,4813 & -0,2980 & 0,4234 \\ 0,8813 & 0,1351 & -0,0160 & -0,1966 & -0,4076 \end{bmatrix},$$

$$A = \begin{bmatrix} 0,1376 & 0,4086 & 0,1594 & 0,4390 & 0,4113 \\ 0,9665 & 0,6246 & 0,3383 & 0,7221 & 0,8746 \\ 0,8285 & 0,0661 & 0,9112 & 0,6266 & 0,2327 \\ 0,0728 & 0,3485 & 0,8560 & 0,8348 & 0,2474 \\ 0,5500 & 0,9198 & 0,0080 & 0,7610 & 1,0506 \end{bmatrix},$$

$$K = \begin{bmatrix} -0,9660 & 0,0 & 0,0 & 0,0569 & -0,2520 \\ 0,1279 & -0,6389 & 0,0 & -0,4688 & -0,5963 \\ -0,0709 & 0,1861 & -0,8383 & -0,4806 & 0,1632 \\ 0,0726 & -0,4109 & -0,5354 & 0,7253 & -0,1144 \\ -0,2002 & -0,6231 & 0,1027 & -0,1411 & 0,7357 \end{bmatrix},$$

$$R = \begin{bmatrix} 1,4446 & 1,6867 & 1,2530 \\ 0,0 & -1,3389 & -0,1486 \\ 0,0 & 0,0 & 1,1831 \end{bmatrix}.$$

Случай 2б. Переопределенная матрица полного ранга:

$$m = 6, n = 5, \text{rank } A = 3.$$

(3.26)

$$QAK = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} -0,0894 & -0,6279 & -0,5382 & -0,0473 & -0,3573 & -0,4221 \\ 0,3768 & 0,0510 & -0,4986 & 0,3522 & 0,6813 & -0,1360 \\ -0,1235 & 0,2301 & -0,3633 & -0,8609 & 0,2386 & 0,0417 \\ 0,3354 & -0,6906 & 0,4092 & -0,2991 & 0,3539 & 0,1685 \\ 0,7849 & 0,1389 & -0,1992 & -0,1088 & -0,4751 & 0,2956 \\ -0,3259 & -0,2324 & -0,3500 & 0,1768 & 0,0153 & 0,8281 \end{bmatrix},$$

$$A = \begin{bmatrix} 0,1376 & 0,4087 & 0,1593 & 0,4308 & 0,4163 \\ 0,9667 & 0,6246 & 0,3384 & 0,8397 & 0,8029 \\ 0,8286 & 0,0661 & 0,9111 & 0,7495 & 0,1577 \\ 0,0728 & 0,3485 & 0,8560 & 0,8068 & 0,2644 \\ 0,5501 & 0,9198 & 0,0080 & 0,7910 & 1,0323 \\ 0,6498 & 0,2725 & 0,3599 & 0,5350 & 0,3801 \end{bmatrix},$$

$$K = \begin{bmatrix} -0,9846 & 0,0 & 0,0 & -0,0868 & -0,1520 \\ 0,1343 & -0,6392 & 0,0 & -0,4230 & -0,6281 \\ 0,0177 & 0,1650 & -0,8460 & -0,4808 & 0,1598 \\ -0,0467 & -0,3820 & -0,5248 & 0,7488 & -0,1256 \\ -0,1006 & -0,6468 & 0,0942 & -0,1469 & 0,7356 \end{bmatrix},$$

$$R = \begin{bmatrix} 1,5636 & 1,7748 & 1,4574 \\ 0,0 & -1,3537 & -0,1233 \\ 0,0 & 0,0 & 1,1736 \end{bmatrix}.$$

Случай 3б. Недоопределенная матрица полного ранга:

$$m = 4, n = 5, \text{rank } A = 3.$$

(3.27)

$$QAK = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} -0,9757 & 0,0103 & 0,1390 & -0,1693 \\ 0,0 & -0,9989 & 0,0298 & -0,0363 \\ 0,0 & 0,0 & -0,7729 & -0,6345 \\ -0,2193 & -0,0458 & -0,6184 & 0,7533 \end{bmatrix},$$

$$A = \begin{bmatrix} 0,4087 & 0,1593 & 0,6594 & 0,4302 & 0,3516 \\ 0,6246 & 0,3383 & 0,6591 & 0,9342 & 0,9038 \\ 0,0661 & 0,9112 & 0,6898 & 0,1931 & 0,1498 \\ 0,2112 & 0,8150 & 0,7983 & 0,3406 & 0,2803 \end{bmatrix},$$

$$K = \begin{bmatrix} -0,4225 & 0,0249 & 0,3156 & -0,0047 & -0,8493 \\ -0,1647 & -0,1529 & -0,9349 & -0,0530 & -0,2696 \\ -0,6816 & 0,6377 & -0,0851 & 0,1253 & 0,3254 \\ -0,4447 & -0,4483 & 0,1187 & -0,7222 & 0,2561 \\ -0,3634 & -0,6070 & 0,0713 & 0,6782 & 0,1858 \end{bmatrix},$$

$$R = \begin{bmatrix} 0,9915 & 0,0 & 0,0 \\ 1,5246 & 0,5840 & 0,0 \\ 1,2573 & -0,1388 & 1,1077 \end{bmatrix}.$$

Упражнения

3.28. Найти спектральное разложение матрицы Хаусхолдера  $H = I - 2ww^T$ ,  $\|w\| = 1$ .

3.29. Найти спектральное разложение матрицы отражения Гивенса из (3.10).

3.30. Показать, что  $G$  из (3.10) является матрицей Хаусхолдера.

## ГЛАВА 4

### ОРТОГОНАЛЬНОЕ РАЗЛОЖЕНИЕ ПОСРЕДСТВОМ СИНГУЛЯРНОГО РАЗЛОЖЕНИЯ

В этой главе будет описано еще одно практически полезное ортогональное разложение  $m \times n$ -матрицы  $A$ . В предыдущей главе матрица  $A$  была представлена произведением  $HRK^T$ , где  $R$  — некоторая прямоугольная матрица, ненулевые элементы которой сосредоточены в невырожденной треугольной подматрице. Мы покажем здесь, что эту невырожденную подматрицу  $R$  можно еще более упростить так, чтобы она стала невырожденной диагональной матрицей. Получаемое в результате разложение особенно полезно при анализе влияния ошибок входной информации на решение задачи НК.

Это разложение тесно связано со спектральным разложением симметричных неотрицательно определенных матриц  $A^T A$  и  $AA^T$ . Стандартные факты, касающиеся спектральных разложений симметричных неотрицательно определенных матриц, приведены в приложении А.

**Теорема 4.1 (сингулярное разложение).** Пусть  $A$  —  $m \times n$ -матрица ранга  $k$ . Тогда существуют ортогональная  $m \times m$ -матрица  $U$ , ортогональная  $n \times n$ -матрица  $V$  и диагональная  $m \times n$ -матрица  $S^*$  такие, что

$$U^T A V = S, \quad A = USV^T. \quad (4.2)$$

Матрицу  $S$  можно выбрать так, чтобы ее диагональные элементы составляли невозрастающую последовательность; все эти элементы неотрицательны и ровно  $k$  из них строго положительны.

Диагональные элементы  $S$  называются сингулярными числами  $A$ . Будет удобно дать доказательство теоремы 4.1 вначале для специального случая

\*) То есть  $m \times n$ -матрица  $S$  такая, что  $s_{ij} \neq 0 \Rightarrow i = j$ . (Примеч. пер.)

$m = n = \text{rank } A$ . Более общее утверждение легко следует из этого специального случая.

**Лемма 4.3.** Пусть  $A$  —  $n \times n$ -матрица ранга  $n$ . Тогда существуют ортогональная  $n \times n$ -матрица  $U$ , ортогональная  $n \times n$ -матрица  $V$  и диагональная  $n \times n$ -матрица  $S$  такие, что

$$U^T A V = S, \quad A = USV^T \quad (4.4)$$

и последовательные диагональные элементы  $S$  положительны и не возрастают.

**Доказательство.** Положительно определенная симметричная матрица  $A^T A$  допускает спектральное разложение

$$A^T A = V D V^T, \quad (4.5)$$

где  $V$  — ортогональная  $n \times n$ -матрица, а  $D$  — диагональная матрица, причем диагональные элементы  $D$  положительны и не возрастают.

Определим  $S$  как диагональную  $n \times n$ -матрицу, диагональные элементы которой суть положительные квадратные корни из соответствующих диагональных элементов  $D$ . Таким образом,

$$D = S^T S = S^2, \quad (4.6)$$

$$S^{-1} D S^{-1} = I_n. \quad (4.7)$$

Определим  $n \times n$ -матрицу

$$U = A V S^{-1}. \quad (4.8)$$

Из (4.5), (4.7), (4.8) и ортогональности  $V$  следует, что

$$U^T U = S^{-1} V^T A^T A V S^{-1} = S^{-1} D S^{-1} = I_n, \quad (4.9)$$

т.е.  $U$  ортогональна.

Из (4.8) и того обстоятельства, что  $V$  ортогональна, выводим

$$USV^T = A V S^{-1} S V^T = A V V^T = A. \quad (4.10)$$

Лемма 4.3 доказана.

**Доказательство теоремы 4.1.** Пусть

$$A = H R K^T, \quad (4.11)$$

где  $H, R, K^T$  имеют свойства, указанные в теореме 3.19.

Так как  $k \times k$ -матрица  $R_{11}$  из (3.21) невырождена, то согласно лемме 4.3, можно написать

$$R_{11} = \tilde{U} \tilde{S} \tilde{V}^T. \quad (4.12)$$

Здесь  $\tilde{U}$  и  $\tilde{V}$  — ортогональные  $k \times k$ -матрицы, а  $\tilde{S}$  — невырожденная диагональная матрица, диагональные элементы которой положительны и не возрастают.

Из (4.12) следует, что матрицу  $R$  уравнения (3.21) можно записать в виде

$$R = \hat{U} \hat{S} \hat{V}^T, \quad (4.13)$$

где  $\hat{U}$  — ортогональная  $m \times m$ -матрица:

$$\hat{U} = \begin{bmatrix} \tilde{U} & 0 \\ 0 & I_{m-k} \end{bmatrix}, \quad (4.14)$$

$\hat{V}$  — ортогональная  $n \times n$ -матрица:

$$\hat{V} = \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_{n-k} \end{bmatrix} \quad (4.15)$$

и  $S$  — диагональная  $m \times n$ -матрица:

$$S = \begin{bmatrix} \tilde{S} & 0 \\ 0 & 0 \end{bmatrix}. \quad (4.16)$$

Теперь, определяя  $U$  и  $V$  формулами

$$U = H\hat{U}, \quad (4.17)$$

$$V = K\hat{V}, \quad (4.18)$$

заключаем из (4.11)–(4.18), что

$$A = USV^T, \quad (4.19)$$

где  $U$ ,  $S$  и  $V$  имеют свойства, указанные в формулировке теоремы 4.1. Это завершает доказательство.

Заметим, что сингулярные числа матрицы  $A$  определены однозначно несмотря на то, что в выборе ортогональных матриц  $U$  и  $V$  из (4.19) есть произвол. Пусть  $\sigma$  — сингулярное число  $A$ , имеющее кратность  $l$ . Это значит, что для упорядоченных сингулярных чисел найдется индекс  $i$  такой, что

$$\sigma = s_j, \quad j = i, i+1, \dots, i+l-1,$$

$$\sigma \neq s_j, \quad j < i \text{ или } j \geq i+l.$$

В  $n$ -мерном пространстве  $l$ -мерное подпространство  $T$ , натянутое на столбцы  $v_j$ ,  $j = i, \dots, i+l-1$ , матрицы  $V$ , определено однозначно; однако однозначности нет при выборе ортогонального базиса в  $T$ , каковым являются названные столбцы  $V$ .

Более точно, положим  $k = \min(m, n)$ , и пусть  $Q$  — ортогональная  $k \times k$ -матрица вида

$$Q = \begin{bmatrix} I_{i-1} & 0 & 0 \\ 0 & P & 0 \\ 0 & 0 & I_{k-i+1} \end{bmatrix}$$

Здесь  $P$  — ортогональная  $l \times l$ -матрица. Если  $A = USV^T$  — сингулярное разложение  $A$  и  $s_i = \dots = s_{i+l-1}$ , то  $\tilde{U}S\tilde{V}^T$ , где

$$\tilde{U} = U \begin{bmatrix} Q & 0 \\ 0 & I_{m-k} \end{bmatrix},$$

$$\tilde{V} = V \begin{bmatrix} Q & 0 \\ 0 & I_{n-k} \end{bmatrix},$$

также будет сингулярным разложением  $A$ .

В следующем численном примере дано сингулярное разложение матрицы  $A$  вида (3.23):

$$A = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T, \quad (4.20)$$

$$U = \begin{bmatrix} -0,4347 & 0,6141 & -0,6587 \\ -0,5509 & -0,7599 & -0,3450 \\ -0,7125 & 0,2129 & 0,6687 \end{bmatrix},$$

$$S = \begin{bmatrix} 0,9965 & 0,0 \\ 0,0 & 0,1128 \end{bmatrix},$$

$$V = \begin{bmatrix} -0,8626 & 0,5058 \\ -0,5058 & -0,8626 \end{bmatrix}.$$

#### Упражнения

4.21. Если  $A$  — симметричная  $n \times n$ -матрица и ее сингулярные числа различны, то:

- а) (действительные) собственные значения  $A$  различны;
- б) с помощью сингулярного разложения матрицы  $A$  можно найти ее спектральное разложение, и наоборот.

Что можно сказать о случае, когда среди сингулярных чисел имеются кратные?

4.22. Если  $s_1$  — наибольшее сингулярное число  $A$ , то  $\|A\| = s_1$ .

4.23. Если  $R$  — невырожденная  $n \times n$ -матрица и  $s_n$  — ее наименьшее сингулярное число, то  $\|R^{-1}\| = s_n^{-1}$ .

4.24. Пусть  $s_1$  и  $s_n$  — соответственно наибольшее и наименьшее сингулярные числа  $m \times n$ -матрицы  $A$  ранга  $n$ . Показать, что для любого  $n$ -вектора  $x$

$$s_n \|x\| \leq \|Ax\| \leq s_1 \|x\|.$$

4.25. Пусть  $s_1, \dots, s_k$  — ненулевые сингулярные числа  $A$ . Тогда

$$\|A\|_F = \left( \sum_{i=1}^k s_i^2 \right)^{1/2}.$$

4.26. Пусть  $A = USV^T$  — сингулярное разложение  $A$ . Показать, что столбцы  $U$  суть собственные векторы симметричной матрицы  $AA^T$ .

## ГЛАВА 5

### ТЕОРЕМЫ О ВОЗМУЩЕНИЯХ СИНГУЛЯРНЫХ ЧИСЕЛ

Сингулярные числа матрицы очень устойчивы к изменению ее элементов. Возмущения элементов матрицы приводят к возмущениям той же или меньшей величины в ее сингулярных числах. Цель данной главы — представить теоремы 5.7 и 5.10, которые дают точные формулировки этой устойчивости, и теорему 5.12, указывающую оценки возмущений сингулярных чисел при удалении из матрицы столбца или строки.

Эти теоремы являются прямыми следствиями соответствующих теорем об устойчивости собственных значений симметричной матрицы. Вначале мы сформулируем три относящиеся сюда теоремы о собственных значениях.

**Теорема 5.1.** Пусть  $B, A, E$  — симметричные  $n \times n$ -матрицы и  $B - A = E$ . Обозначим собственные значения этих матриц, упорядоченные по не-

возрастанию, соответственно через  $\beta_i, \alpha_i, \epsilon_i, i = 1, \dots, n$ . Тогда

$$\epsilon_n \leq \beta_1 - \alpha_1 \leq \epsilon_1, \quad i = 1, \dots, n.$$

Часто бывает полезно более слабое неравенство, которое требует зато менее детальной информации относительно  $E$ :

$$|\beta_i - \alpha_i| \leq \max_j |\epsilon_j| = \|E\|, \quad i = 1, \dots, n.$$

**Теорема 5.2 (Виландта – Хофмана).** В условиях теоремы 5.1

$$\sum_{i=1}^n (\beta_i - \alpha_i)^2 \leq \sum_{i=1}^n \epsilon_i^2 \equiv \sum_{i,j=1}^n e_{ij}^2 \equiv \|E\|_F^2.$$

**Теорема 5.3.** Пусть  $A$  – симметричная  $n \times n$ -матрица с собственными значениями  $\alpha_1 \geq \dots \geq \alpha_n$ . Пусть  $k$  – целое число,  $1 \leq k \leq n$ . Пусть  $B$  – симметричная матрица порядка  $n-1$ , полученная удалением из  $A$   $k$ -й строки и  $k$ -го столбца. Тогда собственные значения  $\beta_i$  матрицы  $B$ , также занумерованные в порядке невозрастания, разделяют собственные значения  $A$ :

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \dots \geq \beta_{n-1} \geq \alpha_n.$$

Обсуждение и доказательства этих трех теорем, равно как и минимаксной теоремы Куранта–Фишера, которую используют теоремы 5.1 и 5.3, читатель может найти в книге [7]. Теорема 5.2 впервые опубликована в [100]; ее обобщение обсуждается в [196].

Чтобы вывести из этих теорем о возмущениях собственных значений теоремы о возмущениях сингулярных чисел, мы воспользуемся связью, существующей между сингулярным разложением матрицы  $A$  и спектральным разложением симметричной матрицы

$$C = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \quad (5.4)$$

Если  $A$  – квадратная и имеет сингулярное разложение  $A = USV^T$ , то легко проверить, что  $C$  имеет спектральное разложение

$$C = \begin{bmatrix} \tilde{U} & -\tilde{U} \\ \tilde{V} & \tilde{V} \end{bmatrix} \cdot \begin{bmatrix} S & 0 \\ 0 & -S \end{bmatrix} \cdot \begin{bmatrix} \tilde{U}^T & \tilde{V}^T \\ -\tilde{U}^T & \tilde{V}^T \end{bmatrix}, \quad (5.5)$$

где  $\tilde{U} = 2^{-1/2}U$  и  $\tilde{V} = 2^{-1/2}V$ .

Если  $m \times n$ -матрица  $A$  ( $m \geq n$ ) имеет сингулярное разложение

$$A = [U_m^{(1)} : U_m^{(2)}] \cdot \begin{bmatrix} S_{n \times n} \\ 0_{(m-n) \times n} \end{bmatrix} \cdot V_{n \times n}^T,$$

то матрица  $C$ , определенная формулой (5.4), имеет спектральное разложение

$$A = P \begin{bmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{bmatrix} P^T,$$

где

$$P = \begin{bmatrix} \tilde{U}^{(1)} & -\tilde{U}^{(1)} & U^{(2)} \\ \tilde{V} & \tilde{V} & 0 \end{bmatrix},$$

$$\tilde{U}^{(1)} = 2^{-1/2}U^{(1)}, \quad \tilde{V} = 2^{-1/2}V.$$

Ясно, что аналогичные результаты справедливы в случае  $m < n$ . Для последующих ссылок мы сформулируем теорему 5.6, вытекающую из вышесказанного.

**Теорема 5.6.** Пусть  $A$  –  $m \times n$ -матрица и  $k = \min(m, n)$ . Пусть  $C$  – симметричная матрица порядка  $m+n$ , определенная формулой (5.4). Если  $s_1, \dots, s_k$  – сингулярные числа  $A$ , то собственные значения  $C$  суть  $s_1, \dots, s_k, -s_1, \dots, -s_k$  и нуль, повторенный  $|m-n|$  раз.

Мы можем теперь сформулировать три теоремы, относящиеся к возмущениям сингулярных чисел.

**Теорема 5.7.** Пусть  $B, A, E$  –  $m \times n$ -матрицы и  $B - A = E$ . Обозначим сингулярные числа этих матриц, упорядоченные по невозрастанию, соответственно через  $\beta_i, \alpha_i, \epsilon_i, i = 1, \dots, k; k = \min(m, n)$ . Тогда

$$|\beta_i - \alpha_i| \leq \epsilon_i \equiv \|E\|, \quad i = 1, \dots, k. \quad (5.8)$$

**Доказательство.** Введем симметричные матрицы

$$\tilde{B} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}, \quad \tilde{E} = \begin{bmatrix} 0 & E \\ E^T & 0 \end{bmatrix}. \quad (5.9)$$

Тогда  $\tilde{B} - \tilde{A} = \tilde{E}$ . Связь собственных значений этих матриц с сингулярными числами  $B, A$  и  $E$  указана в теореме 5.6. Применяя к матрицам  $\tilde{B}, \tilde{A}$  и  $\tilde{E}$  теорему 5.1, получим неравенства 5.8.

**Теорема 5.10.** В условиях теоремы 5.7 справедливо неравенство

$$\sum_{i=1}^k (\beta_i - \alpha_i)^2 \leq \sum_{i=1}^k \epsilon_i^2 \equiv \sum_{i=1}^m \sum_{j=1}^n e_{ij}^2 \equiv \|E\|_F^2. \quad (5.11)$$

**Доказательство.** Вводя те же матрицы (5.9) и используя теоремы 5.6 и 5.2, получим

$$2 \sum_{i=1}^k (\beta_i - \alpha_i)^2 \leq \|\tilde{E}\|_F^2 = 2 \|E\|_F^2,$$

что эквивалентно неравенству (5.11).

**Теорема 5.12.** Пусть  $A$  –  $m \times n$ -матрица. Пусть  $k$  – целое число,  $1 \leq k \leq n$ . Пусть  $B$  –  $m \times (n-1)$ -матрица, полученная из  $A$  удалением  $k$ -го столбца. Если занумеровать сингулярные числа  $\beta_i$  матрицы  $B$  в порядке невозрастания, то они будут разделять сингулярные числа  $\alpha_i$  матрицы  $A$  следующим образом:

1)  $m \geq n$ :

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \dots \geq \beta_{n-1} \geq \alpha_n \geq 0; \quad (5.13)$$

2)  $m < n$ :

$$\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \beta_2 \geq \dots \geq \alpha_m \geq \beta_m \geq 0. \quad (5.14)$$



**Доказательство.** Неравенства (5.13) и (5.14) получаются прямым применением теоремы 5.3 к симметричным матрицам  $\hat{A} = A^T A$  и  $\hat{B} = B^T B$ . В случае 1) собственными значениями  $\hat{A}$  и  $\hat{B}$  будут соответственно числа  $\alpha_i^2$ ,  $i = 1, \dots, n$ , и  $\beta_i^2$ ,  $i = 1, \dots, n-1$ . В случае 2) собственные значения  $\hat{A}$  суть  $\alpha_i^2$ ,  $i = 1, \dots, m$ , и нуль, повторенный  $n-m$  раз, а собственные значения  $\hat{B}$  —  $\beta_i^2$ ,  $i = 1, \dots, m$ , и нуль, повторенный  $n-1-m$  раз.

#### Упражнения

5.15. Задача [51]. Даны  $m \times n$ -матрица  $A$  ранга  $k$  и неотрицательное целое число  $r$ ,  $r < k$ . Найти  $m \times n$ -матрицу  $B$  ранга  $r$ , минимизирующую  $\|B - A\|_F$ .

**Решение.** Пусть  $A = USV^T$  — сингулярное разложение  $A$  с упорядоченными сингулярными числами  $s_1 \geq \dots \geq s_k > 0$ . Пусть  $\tilde{S}$  получена из  $S$  заменой чисел  $s_{r+1}, \dots, s_k$  нулями. Показать, что матрица  $\tilde{B} = U\tilde{S}V^T$  решает поставленную задачу, и выразить  $\|\tilde{B} - A\|_F$  через сингулярные числа  $A$ .

**Замечание.** Это доказательство прямо вытекает из теоремы 5.10. Доказательство, предложенное Экартом и Янгом, по существу, содержит в себе и доказательство теоремы 5.10.

5.16. Решить задачу 5.15, заменив  $\|\cdot\|_F$  на  $\|\cdot\|$ .

5.17. Определим  $\kappa(A)$  как отношение наибольшего сингулярного числа матрицы  $A$  к ее наименьшему ненулевому сингулярному числу. (Это "число обусловленности"  $A$  будет использоваться в последующих главах.) Показать, что если ранг  $m \times n$ -матрицы  $A$  равен  $n$ , а  $m \times n$ -матрица  $B$  получена из  $A$  удалением  $n-r$  столбцов, то  $\kappa(B) < \kappa(A)$ .

## ГЛАВА 6

### ОЦЕНКИ ДЛЯ ЧИСЛА ОБУСЛОВЛЕННОСТИ ТРЕУГОЛЬНОЙ МАТРИЦЫ

При практическом анализе задачи наименьших квадратов в гл. 25 возникнет необходимость в оценке наибольшего и наименьшего ненулевого сингулярных чисел матрицы  $A$ : отношение этих двух величин (число обусловленности  $A$ ) интерпретируется как коэффициент увеличения ошибки. Эта интерпретация числа обусловленности будет представлена в гл. 9.

Наиболее прямой подход состоит в том, чтобы вычислить сингулярные числа  $A$  (см. гл. 4, 18). Однако часто бывает желательно получить оценку для числа обусловленности, не вычисляя сингулярных чисел. В данной главе будут изложены некоторые теоремы и примеры, связанные с этой задачей.

Большинство алгоритмов, описываемых в этой книге, порождают в качестве промежуточного результата невырожденную треугольную матрицу (назовем ее  $R$ ), имеющую те же ненулевые сингулярные числа, что и исходная матрица  $A$ . В общем случае невырожденная треугольная матрица является лучшим отправным пунктом при оценке обусловленности, чем заполненная матрица. Поэтому мы ограничимся оценками сингулярных чисел только для невырожденной треугольной матрицы  $R$ .

Обозначим упорядоченные сингулярные числа невырожденной треугольной  $n \times n$ -матрицы  $R$  через  $s_1 \geq \dots \geq s_n > 0$ . Как отмечено в упражнении 4.22,  $s_1 = \|R\|$ . Отсюда получаем легко вычисляемую нижнюю оценку

для  $s_1$ :

$$s_1 \geq \max_{i,j} |r_{ij}|. \quad (6.1)$$

Далее, так как  $R$  треугольная, то числа, обратные к диагональным элементам  $R$ , суть элементы  $R^{-1}$ . Поэтому

$$\|R^{-1}\| \geq \max_i |r_{ii}^{-1}|. \quad (6.2)$$

Согласно упражнению 4.23,  $s_n^{-1} = \|R^{-1}\|$ , откуда

$$s_n \leq \min_i |r_{ii}|. \quad (6.3)$$

Итак, нижнюю границу  $\rho$  для числа обусловленности  $\kappa = s_1/s_n$  дает неравенство

$$\kappa \geq \rho \equiv \frac{\max_{i,j} |r_{ij}|}{\min_i |r_{ii}|}. \quad (6.4)$$

Эта нижняя граница, хотя и не лишена практической пользы, не может, вообще говоря, рассматриваться как надежная оценка для  $\kappa$ . В самом деле,  $\kappa$  может быть значительно больше, чем  $\rho$ .

Это показывает пример (см. [108]) верхней треугольной  $n \times n$ -матрицы  $R$ , определяемой формулами

$$r_{ij} = \begin{cases} 0, & j < i, \\ -1, & j = i, \\ -1, & j > i. \end{cases} \quad (6.5)$$

Например, для  $n = 4$

$$R = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (6.6)$$

Используя (6.4), получим  $\rho = 1$  в качестве нижней границы для  $\kappa$ . Эта оценка не содержит никакой информации. Для данной матрицы более реалистичскую верхнюю границу для  $s_n$  можно вывести, рассматривая действие  $R$  на вектор  $y$  с компонентами

$$y_i = 2^{1-i}, \quad i = 1, \dots, n-1, \quad y_n = y_{n-1}. \quad (6.7)$$

Легко проверить, что  $Ry = z$ , где

$$z_i = 0, \quad i = 1, \dots, n-1, \quad z_n = 2^{2-n}. \quad (6.8)$$

Теперь, используя неравенства

$$s_n^{-1} = \|R^{-1}\| \geq \frac{\|y\|}{\|z\|} \geq \frac{1}{2^{2-n}}, \quad (6.9)$$

находим

$$\kappa = s_1/s_n \geq 2^{n-2}. \quad (6.10)$$

Таблица 6.1

Последний диагональный элемент  $\tilde{r}_{nn}$  и последнее сингулярное число  $R$ 

| $n$ | $\tilde{r}_{nn} \times 10^8$ | $s_n \times 10^8$ | $s_n/\tilde{r}_{nn}$ |
|-----|------------------------------|-------------------|----------------------|
| 20  | 330                          | 286               | 0,867                |
| 21  | 165                          | 143               | 0,867                |
| 22  | 82,5                         | 71,6              | 0,868                |
| 23  | 40,5                         | 35,8              | 0,884                |
| 24  | 20,5                         | 17,9              | 0,873                |
| 25  | 8,96                         | 8,95              | 0,999                |
| 26  | 4,59                         | 4,52              | 0,985                |

Неравенство (6.9) можно интерпретировать и так: оно показывает, что  $R$  близка к вырожденной матрице. Близость понимается в том смысле, что существует матрица  $E$  такая, что  $\|E\| \leq 2^{2-n}$  и матрица  $R - E$  вырождена. Легко проверить, что нужная матрица задается формулой  $E = zy^T/y^T y$ .

Для  $n = 4$   $y = (1, 1/2, 1/4, 1/4)^T$ ,  $z = (0, 0, 0, 1/4)^T$ ,  $s_4 \leq 1/4$ ,  $\kappa \geq 4$ . Кроме того, при вычитании матрицы

$$E = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{2}{11} & \frac{1}{11} & \frac{1}{22} & \frac{1}{22} \end{bmatrix}$$

из матрицы  $R$  (см. (6.6)) получим вырожденную матрицу.

Этот пример показывает, что иногда опасно считать матрицу хорошо обусловленной даже в том случае, когда  $\rho$  в (6.4) мало, а матрица выглядит вполне "невинно".

Напомним, что нас интересуют главным образом треугольные матрицы, возникающие в алгоритмах решения линейных систем. Матрица, определяемая формулами (6.5), обсуждалась в литературе в связи с тем, что она инвариантна относительно гауссова исключения как с частичным, так и с полным выбором главного элемента. Однако она не будет инвариантна относительно исключения посредством преобразований Хаусхолдера, сопровождаемого перестановками столбцов (так обстоит дело, например, в алгоритме HFTI, который будет описан в гл. 14).

В этой связи и в качестве численного эксперимента мы применим алгоритм HFTI к матрице  $R$  из (6.5). Пусть  $\tilde{R}$  обозначает треугольную матрицу, полученную в результате этой операции. Мы вычислили также сингулярные числа  $R$ . Вычисления проводились на машине UNIVAC 1108 в режиме смешанной точности (см. гл. 17), характеризуемом использованием как арифметики с обычной точностью ( $\eta = 2^{-27} \approx 0,745 \times 10^{-8}$ ), так и арифметики с удвоенной точностью ( $\omega = 2^{-58} \approx 0,347 \times 10^{-17}$ ). В табл. 6.1 представлены вычисленные значения последнего диагонального элемента  $\tilde{r}_{nn}$  и наименьшего сингулярного числа  $s_n$ , а также отношения  $s_n/\tilde{r}_{nn}$  для значе-

ний  $n$  от 20 до 26. Заметим, что во всех случаях  $\tilde{r}_{nn}$  дает хорошую оценку величины  $s_n$ .

Тем не менее существуют  $n \times n$ -матрицы с нормированными столбцами, которые могут быть порождены алгоритмом HFTI и у которых наименьшее сингулярное число приблизительно в  $2^{n-1}$  раз меньше, чем наименьший диагональный элемент.

Пример такой матрицы, который мы сейчас приведем, принадлежит Кахану [108]. Определим  $n \times n$ -матрицу  $R$  формулами

$$r_{ij} = \begin{cases} 0, & i > j, \\ s^{i-1}, & i = j, \\ -cs^{i-1}, & i < j. \end{cases}$$

Здесь  $s$  и  $c$  — положительные числа, причем  $s^2 + c^2 = 1$ . Например, для  $n = 4$

$$R = \begin{bmatrix} 1 & -c & -c & -c \\ 0 & s & -cs & -cs \\ 0 & 0 & s^2 & -cs^2 \\ 0 & 0 & 0 & s^3 \end{bmatrix}. \quad (6.11)$$

Для любого  $n$  матрица  $R$  верхняя треугольная, ее столбцы имеют единичную евклидову длину, и выполняются неравенства

$$r_{kk}^2 \geq \sum_{i=k}^j r_{ij}^2, \quad k = 1, \dots, n-1, \quad j = k+1, \dots, n. \quad (6.12)$$

Как устанавливается в гл. 14, эти неравенства означают, что матрица  $R$  может быть получена в результате применения алгоритма HFTI к некоторой матрице  $A$ .

Пусть  $T = R^{-1}$ . Элементы  $T$  выражаются формулами

$$t_{ij} = \begin{cases} 0, & j < i, \\ \frac{1}{s^{j-1}}, & j = i, \\ \frac{c}{s^{j-1}}, & j = i+1, \\ \frac{c(1+c)^{j-i-1}}{s^{j-1}}, & j-i \geq 2. \end{cases}$$

Например, в случае  $n = 4$

$$R^{-1} = T = \begin{bmatrix} 1 & \frac{c}{s} & \frac{c(1+c)}{s^2} & \frac{c(1+c)^2}{s^3} \\ 0 & \frac{1}{s} & \frac{c}{s^2} & \frac{c(1+c)}{s^3} \\ 0 & 0 & \frac{1}{s^2} & \frac{c}{s^3} \\ 0 & 0 & 0 & \frac{1}{s^3} \end{bmatrix}$$

Пусть  $\tau_n$  обозначает последний столбец  $R^{-1}$ . Тогда

$$\|\tau_n\|^2 = \left(1 + \frac{c^2[(1+c)^{2n-2} - 1]}{(1+c)^2 - 1}\right) r_{nn}^{-2}.$$

Когда  $s \rightarrow +0$ , а  $c = (1-s^2)^{1/2} \rightarrow 1-0$ , произведение  $\|\tau_n\|^2 r_{nn}^2$ , возрастающая, стремится к  $(4^{n-1} + 2)/3$ . Поэтому найдется значение  $s$ , зависящее от  $n$ , для которого

$$\|\tau_n\|^2 r_{nn}^2 \geq 4^{n-1}/3.$$

Так как, согласно упражнению 4.24,

$$s_n \leq \frac{\|R\tau_n\|}{\|\tau_n\|} = \frac{1}{\|\tau_n\|},$$

то

$$s_n \leq 3^{1/2} 2^{1-n} |r_{nn}| \approx 1.73 \cdot 2^{1-n} |r_{nn}|.$$

Следующая теорема показывает, что в рассмотренном примере реализуется почти минимальное значение, которого может достигать отношение  $s_n/|r_{nn}|$ .

**Теорема 6.13 [9].** Пусть  $A$  —  $m \times n$ -матрица ранга  $n$ . Предположим, что все столбцы  $A$  имеют единичную евклидову длину. Пусть  $R$  — верхняя треугольная матрица, полученная процессом хаусхолдеровой триангуляции  $A$  с перестановками столбцов (так обстоит дело, например, в алгоритме HFTI из (14.9)). Тогда  $s_n$  — наименьшее сингулярное число  $A$  и  $r_{nn}$  — последний диагональный элемент  $R$  связаны неравенствами

$$s_n \leq |r_{nn}|, \quad (6.14)$$

$$s_n \geq 3(4^n + 6n - 1)^{-1/2} |r_{nn}| \geq 2^{1-n} |r_{nn}|. \quad (6.15)$$

**Доказательство.** Из способа, которым строится  $R$ , видно, что ее столбцы имеют единичную евклидову длину, а сингулярные числа те же, что и у  $A$ . Кроме того, перестановки столбцов обеспечивают выполнение неравенств

$$r_{kk}^2 \geq \sum_{i=k}^j r_{ij}^2, \quad k=1, \dots, n-1, \quad j=k+1, \dots, n. \quad (6.16)$$

Из (6.16) следует, что

$$|r_{kk}| \geq |r_{ij}|, \quad i \geq k, \quad j > k, \quad k=1, \dots, n-1. \quad (6.17)$$

Так как

$$s_n = \|R^{-1}\|^{-1}, \quad (6.18)$$

то неравенства (6.14) и (6.15) эквивалентны неравенствам

$$\|R^{-1}\| \geq |r_{nn}^{-1}|, \quad (6.19)$$

$$\|R^{-1}\| \leq 3^{-1}(4^n + 6n - 1)^{1/2} |r_{nn}^{-1}|. \quad (6.20)$$

Неравенство (6.19) — следствие того факта, что  $r_{nn}^{-1}$  является элементом матрицы  $R^{-1}$ .

Чтобы доказать (6.20), мы выведем верхние границы для величин всех элементов  $R^{-1}$ , а затем вычислим норму Фробениуса мажорирующей матрицы  $\tilde{M}$ .

Положим  $T = R^{-1}$ . Будет удобно сгруппировать элементы  $T$  по диагоналям, параллельным главной. Введем мажорирующие параметры для диагоналей следующим образом:

$$g_k = \max\{|t_{i,i+k}|, i=1, \dots, n-k\}, \quad k=0, 1, \dots, n-1. \quad (6.21)$$

Тогда

$$g_0 = \max_i |t_{ii}| = \max_i |r_{ii}^{-1}| = |r_{nn}^{-1}|. \quad (6.22)$$

Элементы  $k$ -й наддиагонали  $T$  можно выразить через элементы главной диагонали  $T$ , элементы предшествующих наддиагоналей  $T$  и элементы  $R$ :

$$t_{i,i+k} = - \frac{\sum_{l=1}^k r_{i,i+l} t_{i+l,i+k}}{r_{ii}}, \quad (6.23)$$

$$1 \leq i \leq n-1, \quad 1 \leq k \leq n-i.$$

Так как  $|r_{i,i+l}| \leq |r_{ii}|$ , то из (6.21) и (6.23) получаем

$$g_k = \max_i |t_{i,i+k}| \leq \max_i \sum_{l=1}^k |t_{i+l,i+k}| \leq \sum_{l=1}^k g_{k-l} = \sum_{j=0}^{k-1} g_j. \quad (6.24)$$

Легко проверить по индукции, что

$$g_k \leq 2^{k-1} g_0 = 2^{k-1} |r_{nn}^{-1}|, \quad k=1, \dots, n. \quad (6.25)$$

Определим верхнюю треугольную  $n \times n$ -матрицу  $M$  формулами

$$m_{ij} = \begin{cases} 0, & j < i, \\ 1, & j = i, \\ 2^{j-i-1}, & j > i, \end{cases} \quad (6.26)$$

и положим

$$\tilde{M} = |r_{nn}^{-1}| M. \quad (6.27)$$

Например, для  $n=4$

$$M = \begin{bmatrix} 1 & 1 & 2 & 4 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Из (6.21) и (6.25)–(6.27) следует, что элементы  $R^{-1}$  ( $= T$ ) мажорируются соответствующими элементами  $\tilde{M}$ . Поэтому

$$s_n^{-1} = \|R^{-1}\| \leq \|R^{-1}\|_F \leq \|\tilde{M}\|_F = |r_{nn}^{-1}| \|M\|_F. \quad (6.28)$$

Чтобы вычислить норму Фробениуса матрицы  $M$ , заметим, что для  $j \geq 2$

сумма квадратов элементов столбца  $j$  равна

$$w_j^2 = 1 + \sum_{i=0}^{j-2} 4^i = \frac{4^{j-1} + 2}{3}. \quad (6.29)$$

Это выражение верно и при  $j = 1$ , если считать, что сумма в средней части равна нулю.

Теперь

$$\|R^{-1}\|_F^2 = \|T\|_F^2 \leq r_{nn}^{-2} \|M\|_F^2 = r_{nn}^{-2} \sum_{j=1}^n w_j^2 = r_{nn}^{-2} \frac{4^n + 6n - 1}{9}. \quad (6.30)$$

Неравенства (6.28) и (6.30) вместе устанавливают неравенство (6.20). Теорема 6.13 доказана.

**Т е о р е м а 6.31 [9].** Пусть  $A$  и  $R$  те же, что и в теореме 6.13. Сингулярные числа  $s_1 \geq \dots \geq s_n$  матрицы  $A$  связаны с диагональными элементами  $r_{ii}$  матрицы  $R$  неравенствами

$$2^{1-i} |r_{ii}| \leq 3(4^i + 6i - 1)^{-1/2} |r_{ii}| \leq s_i \leq (n - i + 1)^{1/2} |r_{ii}|, \quad i = 1, \dots, n. \quad (6.32)$$

**Д о к а з а т е л ь с т в о.** Пусть  $R_j$  — ведущая главная подматрица порядка  $j$  матрицы  $R$ . Обозначим через  $s_i^{(j)}$   $i$ -е сингулярное число  $R_j$ . Из теоремы 5.12 вытекают неравенства

$$s_i = s_i^{(n)} \geq s_i^{(n-1)} \geq \dots \geq s_i^{(i)}. \quad (6.33)$$

Применяя к  $R_i$  теорему 6.13, получим

$$s_i^{(i)} \geq 3(4^i + 6i - 1)^{-1/2} |r_{ii}|. \quad (6.34)$$

Это неравенство вместе с (6.33) доказывает нижнюю оценку для  $s_i$  в (6.32).

Определим  $W_j$  как главную подматрицу порядка  $n + 1 - j$ , стоящую в  $R$  на пересечении строк и столбцов с номерами  $j, \dots, n$ . Заметим, что первым диагональным элементом  $W_i$  является  $r_{ii}$ . Используя (6.16) и упражнение 6.37, получаем

$$\|W_i\| \leq (n + 1 - i)^{1/2} |r_{ii}|. \quad (6.35)$$

Обозначим через  $\omega_i^{(j)}$   $i$ -е сингулярное число  $W_j$ . Согласно теореме 5.12, можно написать

$$s_i = \omega_i^{(1)} \leq \omega_{i-1}^{(2)} \leq \dots \leq \omega_i^{(i)}. \quad (6.36)$$

Так как  $\omega_i^{(i)} = \|W_i\|$ , то (6.35) и (6.36) вместе устанавливают верхнюю оценку для  $s_i$  в (6.32). Теорема 6.31 доказана.

**У п р а ж н е н и е**

6.37. Пусть  $A$  —  $m \times n$ -матрица со столбцами  $a_j$ . Тогда

$$\|A\| \leq n^{1/2} \max_j \|a_j\|.$$

## ГЛАВА 7

### ПСЕВДООБРАТНАЯ МАТРИЦА

Если  $A$  — невырожденная  $n \times n$ -матрица, то решение системы  $Ax = b$  можно записать в виде  $x = A^{-1}b$ , где  $A^{-1}$  — (единственная) обратная матрица для  $A$ . Обратная матрица — это очень полезное математическое понятие даже при том, что эффективные и надежные современные методы [7, 11] для решения системы  $Ax = b$  не требуют явного вычисления  $A^{-1}$ .

В случае задачи НК возникает вопрос: существует ли  $n \times m$ -матрица  $Z$ , однозначно определяемая матрицей  $A$  и такая, что (единственное) решение минимальной длины задачи НК выражается формулой  $x = Zb$ . Такая матрица  $Z$  действительно существует; она называется *псевдообратной* для матрицы  $A$ . Как и обычная обратная, псевдообратная матрица — полезное математическое понятие, но при решении задачи НК ее, как правило, не вычисляют в явном виде.

Следующие две теоремы приводят к конструктивному определению псевдообратной для  $m \times n$ -матрицы  $A$ .

**Т е о р е м а 7.1.** Пусть  $A$  —  $m \times n$ -матрица ранга  $k$  с ортогональным разложением  $A = HRK^T$ , удовлетворяющим предположениям теоремы 2.3. Тогда единственное решение минимальной длины задачи НК выражается формулой

$$x = K \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} H^T b. \quad (7.2)$$

**Д о к а з а т е л ь с т в о.** Формула (7.2) есть просто иной способ записи равенств (2.5)–(2.7) и (2.11).

**Т е о р е м а 7.3.** Пусть  $A = HRK^T$ , как и в теореме 7.1,

$$Z = K \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} H^T.$$

Матрица  $Z$  однозначно определена матрицей  $A$  и не зависит от конкретного ортогонального разложения  $A$ .

**Д о к а з а т е л ь с т в о.** Для каждого  $j$ ,  $1 \leq j \leq m$ ,  $j$ -й столбец  $z_j$  матрицы  $Z$  можно записать в виде  $z_j = Ze_j$ , где  $e_j$  —  $j$ -й столбец единичной матрицы  $I_m$ . Согласно теореме 7.1,  $z_j$  — единственное решение минимальной длины для задачи наименьших квадратов  $Ax \approx e_j$ . Теорема 7.3 доказана.

Исходя из теорем 7.1 и 7.3, дадим следующее определение.

**О п р е д е л е н и е 7.4.** Для произвольной  $m \times n$ -матрицы  $A$  псевдообратная матрица, обозначаемая через  $A^+$ , — это  $n \times m$ -матрица,  $j$ -й столбец  $z_j$  которой является единственным решением минимальной длины для задачи наименьших квадратов  $Ax_j \approx e_j$ , где  $e_j$  —  $j$ -й столбец единичной матрицы  $I_m$ .

Это определение с учетом теорем 7.1 и 7.3 непосредственно приводит к записи решения минимальной длины задачи НК в виде

$$\tilde{x} = A^+ b. \quad (7.5)$$

Следующие два случая заслуживают специального упоминания. Для квадратной невырожденной матрицы  $B$  псевдообратная матрица совпадает с обычной обратной:

$$B^+ = B^{-1}. \quad (7.6)$$

Для  $m \times n$ -матрицы

$$R = \begin{bmatrix} R_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

где  $R_{11}$  — невырожденная  $k \times k$ -матрица, псевдообратная — это  $n \times m$ -матрица вида

$$R^+ = \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \quad (7.7)$$

Псевдообратную для  $m \times n$ -матрицы  $A$  можно охарактеризовать и другими способами. Каждая из двух следующих теорем дает иную характеристику псевдообратной матрицы.

**Теорема 7.8.** Если  $A = HRK^T$  — произвольное ортогональное разложение  $A$ , удовлетворяющее условиям теоремы 2.3, то  $A^+ = KR^+H^T$ , где  $R^+$  задана формулой (7.7).

**Теорема 7.9** (условия Пенроуза [138]). Псевдообратная  $A^+$  для  $m \times n$ -матрицы  $A$  — это единственная  $n \times m$ -матрица  $X$ , которая удовлетворяет следующей системе условий:

$$a) AXA = A; \quad b) XAX = X; \quad c) (AX)^T = AX; \quad d) (XA)^T = XA.$$

Доказательства этих двух теорем предоставляются читателю в качестве упражнений.

Явное представление  $A^+$ , указанное в теореме 7.8, особенно полезно для вычислений. Если, например, для  $A$  имеется ортогональное разложение (3.20) и  $R_{11}$  в (3.21) — невырожденная треугольная матрица, то

$$A^+ = K \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} H^T. \quad (7.10)$$

Если же имеем сингулярное разложение (4.2), описываемое в теореме 4.1, то

$$A^+ = VS^+U^T. \quad (7.11)$$

Как отмечено в начале этой главы, обычно нет нужды в явном построении псевдообратной для матрицы  $A$ . Например, если целью вычислений является решение задачи НК, то более экономично как по времени, так и по памяти разбить процесс решения на следующие три шага (которые, по существу, повторяют доказательство теоремы 2.3):

$$1) g = H^T b; \quad (7.12)$$

$$2) \text{ решить относительно } y_1 \text{ систему } R_{11} y_1 = g_1; \quad (7.13)$$

$$3) x = K \begin{bmatrix} y_1 \\ 0 \end{bmatrix} = A^+ b. \quad (7.14)$$

## Упражнения

7.15. Проверить формулы (7.6) и (7.7).

7.16. Доказать теорему 7.8.

7.17. [138]. Доказать теорему 7.9.

7.18. Доказать, что  $(A^+)^T = (A^T)^+$  и  $(A^+)^+ = A$ .

7.19. Если  $Q_{m \times n}$  имеет ортонормальные столбцы или ортонормальные строки, то  $Q^+ = Q^T$ . Если  $Q_{m \times n}$  имеет ранг  $n$  и удовлетворяет условию  $Q^+ = Q^T$ , то столбцы  $Q$  ортонормальны.

7.20 [138]. Если  $U$  и  $V$  имеют ортонормальные столбцы, то  $(UAV^T)^+ = VA^+U^T$ .

7.21. а) Пусть  $A = USV^T$  — сингулярное разложение  $A$ . Выписать выражения для сингулярных разложений четырех матриц  $P_1 = A^+A$ ,  $P_2 = I - A^+A$ ,  $P_3 = AA^+$ ,  $P_4 = I - AA^+$  через матрицы  $U, S, V$ .

б) Вывести из а), что матрицы  $P_i$  являются симметричными идемпотентами и, следовательно, проекционными матрицами.

с) Пусть  $T_i$  — подпространство, ассоциированное с проекционной матрицей  $P_i$ , т.е.  $T_i = \{x: P_i x = x\}$ . Указать связь каждого подпространства  $T_i$  с пространством строк или пространством столбцов  $A$ .

7.22. Доказать, что общее решение задачи  $Ax = b$  выражается формулой  $x = A^+ b + (I - A^+ A)u$ , где  $u$  — произвольный вектор.

7.23 [84]. Для невырожденных матриц имеется тождество  $(AB)^{-1} = B^{-1}A^{-1}$ . Аналогичное соотношение  $(AB)^+ = B^+A^+$  выполняется не для всех матриц  $A_{m \times k}$  и  $B_{k \times n}$ .

а) Указать две матрицы  $A$  и  $B$  такие, что  $m = n = 1$ ,  $k = 2$  и  $(AB)^+ \neq B^+A^+$ .

б) Доказать, что матрицы  $A$  и  $B$  удовлетворяют соотношению  $(AB)^+ = B^+A^+$  тогда и только тогда, когда образ  $B$  является инвариантным подпространством для  $A^T A$ , а образ  $A^T$  (т.е. пространство строк  $A$ ) является инвариантным подпространством для  $BB^T$ .

7.24. Если  $\text{rank } A_{m \times n} = n$ , то  $A^+ = (A^T A)^{-1} A^T$ .

Если  $\text{rank } A_{m \times n} = m$ , то  $A^+ = A^T (A A^T)^{-1}$ .

7.25 [81, 138]. Для произвольной матрицы  $A$  уравнения  $XA A^T = A^T$  и  $A^T A Y = A^T$  совместны. Если матрицы  $X$  и  $Y$  суть решения указанных уравнений, то  $XA$  и  $A Y$  — проекционные матрицы (симметричные идемпотенты) и  $A^+ = X A Y$ . Отметьте упрощение формулировки в случае симметричной матрицы  $A$ .

7.26 [81]. Псевдообратная для прямоугольной матрицы  $A$  может быть определена через псевдообратные для симметричных матриц  $A^T A$  или  $A A^T$  (следует выбирать более удобную) формулами  $A^+ = (A^T A)^+ A^T$  или  $A^+ = A^T (A A^T)^+$  соответственно.

7.27 [138]. Если  $A$  нормальная (т.е. удовлетворяет условию  $A^T A = A A^T$ ), то  $A^+ A = A A^+$  и  $(A^n)^+ = (A^+)^n$ .

7.28 [138]. Если  $A = \sum A_i$ , причем  $A_i A_j^T = 0$  и  $A_i^T A_j = 0$  для  $i \neq j$ , то  $A^+ = \sum A_i^+$ .

## ГЛАВА 8

### ОЦЕНКИ ВОЗМУЩЕНИЙ ДЛЯ ПСЕВДООБРАТНЫХ МАТРИЦ

Наша цель здесь и в гл. 9 — изучение связи между возмущениями входных данных задачи НК и возмущениями ее решений. В этой главе будут получены теоремы о возмущениях псевдообратных матриц, которые используются в гл. 9 для исследования возмущений задачи НК.

На практике к рассмотрению таких возмущений может побуждать ограниченная точность, с которой наблюдаемое явление описывается количественной информацией. Влияние погрешностей округлений, производимых в ходе

численной процедуры, тоже можно проанализировать так, как если бы оно имело причиной возмущение входных данных. Подобный анализ для алгоритмов, основанных на преобразованиях Хаусхолдера, будет дан в гл. 15–17.

Результаты, относящиеся к возмущениям псевдообратных матриц или решений задачи НК, были получены рядом авторов. Для наших целей наиболее подходящей с точки зрения как общности, так и удобной формы представления конечных результатов является трактовка, предложенная в [188]. Более ранний анализ проблемы возмущений или ее специальных случаев дан в [23, 24, 80, 90, 139, 171].

Пусть  $A$  и  $E$  —  $m \times n$ -матрицы. Определим возмущенную матрицу

$$\tilde{A} = A + E \quad (8.1)$$

и матрицу-разность

$$G = \tilde{A}^+ - A^+. \quad (8.2)$$

Мы хотим определить зависимость  $G$  от  $E$  и, в частности, получить оценки для  $\|G\|$  через  $\|A\|$  и  $\|E\|$ .

Удобно ввести четыре проекционные матрицы:

$$P = A^+ A = A^T A^T, \quad Q = A A^+ = A^T A^T, \quad (8.3)$$

$$\tilde{P} = \tilde{A}^+ \tilde{A} = \tilde{A}^T \tilde{A}^T, \quad \tilde{Q} = \tilde{A} \tilde{A}^+ = \tilde{A}^T \tilde{A}^T. \quad (8.4)$$

У этих матриц есть ряд полезных свойств, выводимых непосредственно из условий Пенроуза (теорема 7.9). См. также упражнение 7.21 и сводку стандартных свойств проекционных матриц, содержащуюся в приложении А.

Матрицы, определяемые формулами (8.1) — (8.4), будут использоваться на протяжении всей главы без ссылок на эти формулы.

**Теорема 8.5.** Матрица  $G$ , определенная формулами (8.1) и (8.2), может быть представлена в виде

$$G = G_1 + G_2 + G_3, \quad (8.6)$$

где

$$G_1 = -\tilde{A}^+ E A^+, \quad (8.7)$$

$$G_2 = \tilde{A}^+ (I - Q) = \tilde{A}^+ \tilde{A}^T E^T (I - Q), \quad (8.8)$$

$$G_3 = -(I - \tilde{P}) A^+ = (I - \tilde{P}) E^T A^T A^+. \quad (8.9)$$

Для этих матриц справедливы оценки

$$\|G_1\| \leq \|E\| \|A^+\| \|\tilde{A}^+\|, \quad (8.10)$$

$$\|G_2\| \leq \|E\| \|\tilde{A}^+\|^2, \quad (8.11)$$

$$\|G_3\| \leq \|E\| \|A^+\|^2. \quad (8.12)$$

**Доказательство.** Представим  $G$  как сумму следующих восьми матриц:

$$\begin{aligned} G &= [\tilde{P} + (I - \tilde{P})] (\tilde{A}^+ - A^+) [Q + (I - Q)] = \\ &= \tilde{P} \tilde{A}^+ Q + \tilde{P} \tilde{A}^+ (I - Q) - \tilde{P} A^+ Q - \tilde{P} A^+ (I - Q) + \\ &+ (I - \tilde{P}) \tilde{A}^+ Q + (I - \tilde{P}) \tilde{A}^+ (I - Q) - (I - \tilde{P}) A^+ Q + (I - \tilde{P}) A^+ (I - Q). \end{aligned} \quad (8.13)$$

Используя свойства

$$\tilde{P} \tilde{A}^+ = \tilde{A}^+, \quad (I - \tilde{P}) \tilde{A}^+ = 0,$$

$$A^+ Q = A^+, \quad A^+ (I - Q) = 0,$$

приводим (8.13) к виду

$$\begin{aligned} G &= (\tilde{A}^+ Q - \tilde{P} A^+) + \tilde{A}^+ (I - Q) - (I - \tilde{P}) A^+ \equiv \\ &\equiv G_1 + G_2 + G_3. \end{aligned}$$

Чтобы выявить линейную зависимость  $G$  от  $E$ , напомним:

$$\begin{aligned} G_1 &= \tilde{A}^+ A A^+ - \tilde{A}^+ \tilde{A} A^+ = -\tilde{A}^+ E A^+, \\ G_2 &= \tilde{A}^+ \tilde{Q} (I - Q) = \tilde{A}^+ \tilde{A}^T \tilde{A}^T (I - Q) = \\ &= \tilde{A}^+ \tilde{A}^T (\tilde{A}^T - A^T) (I - Q) = \tilde{A}^+ \tilde{A}^T E^T (I - Q), \\ G_3 &= -(I - \tilde{P}) P A^+ = -(I - \tilde{P}) A^T A^T A^+ = \\ &= -(I - \tilde{P}) (A^T - \tilde{A}^T) A^T A^+ = (I - \tilde{P}) E^T A^T A^+. \end{aligned} \quad (8.14)$$

Оценки (8.10)–(8.12) вытекают из неравенств  $\|I - Q\| \leq 1$  и  $\|I - \tilde{P}\| \leq 1$ . Теорема 8.5 доказана.

Заметим, что для действительных чисел  $a$  и  $\tilde{a}$  (а также для квадратных невырожденных матриц) имеется алгебраическое тождество

$$\tilde{a}^{-1} - a^{-1} = a^{-1} (a - \tilde{a}) \tilde{a}^{-1}.$$

Казалось бы, оно означает, что для  $\|G\|$  следовало ожидать оценки вида (8.10). Однако в (8.6) появились добавочные члены  $G_2$  и  $G_3$ . Их присутствие связано с тем, что матрица либо неквадратная, либо квадратная, но невырожденная. Именно:  $G_2$  может быть ненулевой матрицей, только если  $\text{rank } A < m$ , а  $G_3$  может быть ненулевой матрицей, только если  $\text{rank } A < n$ . Если же  $\text{rank } A = m$ , то  $Q = I_m$ , а при  $\text{rank } \tilde{A} = n$  будет  $\tilde{P} = I_n$ .

Теперь мы хотим заменить  $\|\tilde{A}^+\|$  в правых частях неравенств (8.10) и (8.11) ее оценкой через  $\|A^+\|$  и  $\|E\|$ . Получить такую оценку можно при предположениях (8.16) и (8.17) следующей теоремы.

**Теорема 8.15.** Предположим, что

$$\text{rank } (A + E) \leq \text{rank } A = k > 1, \quad (8.16)$$

$$\|A^+\| \|E\| < 1. \quad (8.17)$$

Пусть  $s_k$  — наименьшее ненулевое сингулярное число матрицы  $A$ , а  $\epsilon = \|E\|$ . Тогда

$$\text{rank } (A + E) = k, \quad (8.18)$$

$$\|(A + E)^+\| \leq \frac{\|A^+\|}{1 - \|A^+\| \|E\|} = \frac{1}{s_k - \epsilon}. \quad (8.19)$$

**Доказательство.** Неравенство (8.17) можно переписать в виде  $\epsilon/s_k < 1$  или, что эквивалентно,  $s_k - \epsilon > 0$ . Пусть  $\tilde{s}_k$  —  $k$ -е сингулярное число матрицы  $\tilde{A} \equiv A + E$ . Согласно теореме (5.7),

$$\tilde{s}_k \geq s_k - \epsilon, \quad (8.20)$$

поэтому  $\text{rank } (A + E) \geq k$ , и в силу (8.16) верно (8.18). Неравенство (8.20)

можно записать в виде

$$\frac{1}{\tilde{s}_k} < \frac{1}{s_k - \epsilon}$$

что эквивалентно неравенству (8.19). Теорема 8.15 доказана.

Условия (8.16) и (8.17) необходимы. Легко проверить, что  $\|\tilde{A}^+\|$  может быть неограниченной, если любое из них нарушено. Если же требовать выполнения (8.16) и (8.17), то, помимо оценки (8.19), мы можем, используя условие (8.18), показать, что в оценке (8.11) для  $\|G_2\|$  произведение  $\|E\|\|\tilde{A}^+\|^2$  можно заменить на  $\|E\|\|A^+\|\|\tilde{A}^+\|$ . Это утверждение основывается на теоремах 8.21 и 8.22.

Теорема 8.21. Если  $\text{rank } \tilde{A} = \text{rank } A$ , то

$$\|\tilde{Q}(I - Q)\| = \|Q(I - \tilde{Q})\|.$$

Доказательство. Это доказательство принадлежит Крогу [112].

Запишем сингулярные разложения матриц  $A$  и  $\tilde{A}$ :  $A = USV^T$  и  $\tilde{A} = \tilde{U}\tilde{S}\tilde{V}^T$ . Из (8.3), (8.4) и предположения о том, что  $A$  и  $\tilde{A}$  имеют одинаковый ранг (скажем,  $k$ ), следует

$$Q = U \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} U^T, \quad \tilde{Q} = \tilde{U} \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \tilde{U}^T.$$

Определим ортогональную  $m \times m$ -матрицу  $W$  с элементами  $W_{ij}$  соотношением

$$\tilde{U}^T U = W \equiv \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix}$$

Тогда

$$\begin{aligned} \|\tilde{Q}(I - Q)\| &= \left\| \tilde{U} \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \tilde{U}^T U \begin{bmatrix} 0 & 0 \\ 0 & I_{m-k} \end{bmatrix} U^T \right\| = \\ &= \left\| \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} W \begin{bmatrix} 0 & 0 \\ 0 & I_{m-k} \end{bmatrix} \right\| = \left\| \begin{bmatrix} 0 & W_{12} \\ 0 & 0 \end{bmatrix} \right\| = \|W_{12}\|. \end{aligned}$$

Точно так же можно проверить, что  $\|Q(I - \tilde{Q})\| = \|W_{12}\|$ . Остается показать, что  $\|W_{12}\| = \|W_{21}\|$ . Пусть  $x$  — произвольный  $m - k$ -вектор. Положим

$$y = \begin{bmatrix} 0 \\ x \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix}$$

Используя ортогональность  $W$ , имеем  $\|x\|^2 = \|y\|^2 = \|Wy\|^2 = \|W_{12}x\|^2 + \|W_{22}x\|^2$ , откуда  $\|W_{12}x\|^2 = \|x\|^2 - \|W_{22}x\|^2$ . Следовательно,

$$\|W_{12}\|^2 = \max_{\|x\|=1} \|W_{12}x\|^2 = 1 - \min_{\|x\|=1} \|W_{22}x\|^2 = 1 - s_{m-k}^2,$$

где  $s_{m-k}$  — наименьшее сингулярное число  $W_{22}$ .

Аналогично из  $\|x\|^2 = \|y\|^2 = \|W^T y\|^2 = \|W_{21}^T x\|^2 + \|W_{22}^T x\|^2$  получаем  $\|W_{21}\|^2 = 1 - \min_{\|x\|=1} \|W_{22}^T x\|^2 = 1 - s_{m-k}^2$ .

Итак,  $\|W_{12}\| = \|W_{21}\|$ , что и требовалось доказать.

Теорема 8.22. Если  $\text{rank } \tilde{A} = \text{rank } A$ , то матрица  $G_2$ , определенная в (8.8), удовлетворяет оценке

$$\|G_2\| \leq \|E\|\|A^+\|\|\tilde{A}^+\|. \quad (8.23)$$

Доказательство. Так как  $\text{rank } \tilde{A} = \text{rank } A$ , то теорема 8.21 позволяет написать

$$\|G_2\| \leq \|\tilde{A}^+\| \|\tilde{Q}(I - Q)\| = \|\tilde{A}^+\| \|Q(I - \tilde{Q})\|.$$

Поскольку

$$\begin{aligned} Q(I - \tilde{Q}) &= A^T A^T (I - \tilde{Q}) = \\ &= A^T (A^T - \tilde{A}^T) (I - \tilde{Q}) = -A^T E^T (I - \tilde{Q}), \end{aligned}$$

то  $\|G_2\|$  удовлетворяет неравенству (8.23), что и требовалось доказать.

Теперь мы в состоянии доказать нижеследующую теорему, которую можно рассматривать как наиболее полезный частный случай теоремы 8.5. При более ограничительных предположениях будут получены оценки, не включающие  $\|\tilde{A}^+\|$ .

Теорема 8.24. Пусть  $G_i$ ,  $i = 1, 2, 3$ , — матрицы, определенные соотношениями (8.7)–(8.9). Предположим дополнительно, что  $\|E\|\|A^+\| < 1$  и  $\text{rank } \tilde{A} \leq \text{rank } A$ . Тогда  $\text{rank } \tilde{A} = \text{rank } A$  и

$$\|G_1\| \leq \frac{\|E\|\|A^+\|^2}{1 - \|E\|\|A^+\|}, \quad (8.25)$$

$$\|G_2\| \leq \frac{\|E\|\|A^+\|^2}{1 - \|E\|\|A^+\|}, \quad (8.26)$$

$$\|G_3\| \leq \|E\|\|A^+\|^2, \quad (8.27)$$

$$\|G_4\| \leq \frac{c \|E\|\|A^+\|^2}{1 - \|E\|\|A^+\|}, \quad (8.28)$$

где

$$c = \frac{1 + 5^{1/2}}{2} \approx 1,618, \text{ если } \text{rank } \tilde{A} < \min(m, n), \quad (8.29)$$

$$c = 2^{1/2} \approx 1,414, \text{ если } \text{rank } A = \min(m, n) < \max(m, n), \quad (8.30)$$

$$c = 1, \text{ если } \text{rank } A = m = n. \quad (8.31)$$

Доказательство. То, что  $\text{rank } \tilde{A} = \text{rank } A$ , установлено в теореме 8.15. Используя оценку для  $\|\tilde{A}^+\|$  из этой теоремы, выводим неравенства (8.25), (8.26) и (8.27) из неравенств (8.10), (8.23) и (8.12) соответственно.

Отсюда сразу получается оценка (8.28) с  $c = 3$ . Для практических целей этого результата вполне достаточно. Оценка для случая (8.31) также оче-

видна, поскольку здесь  $\|G_2\| = \|G_3\| = 0$ . Конечно, это неравенство для квадратных невырожденных матриц хорошо известно и может быть доказано непосредственно (см., например, [7]).

Для случаев (8.29) и (8.30) значение  $\alpha$  определяется следующим образом. Пусть  $x$  — нормированный  $m$ -вектор. Положим

$$x_1 = Qx, \quad x_2 = (I - Q)x.$$

Тогда  $x = x_1 + x_2$ ,  $1 = \|x\|^2 = \|x_1\|^2 + \|x_2\|^2$  и найдется число  $\varphi$  такое, что  $\cos \varphi = \|x_1\|$  и  $\sin \varphi = \|x_2\|$ . Пусть  $\alpha \geq \|E\| \|A^+\|^2 / (1 - \|E\| \|A^+\|)$ . Согласно неравенствам (8.25)–(8.27),  $\alpha \geq \|G_i\|$ ,  $i = 1, 2, 3$ . Учитывая правые сомножители  $A^+$  и  $I - Q$  в формулах (8.7)–(8.9), имеем

$$Gx = G_1x_1 + G_2x_2 + G_3x_1 \equiv y_1 + y_2 + y_3.$$

Левые сомножители в (8.7)–(8.9) — это соответственно  $\tilde{A}$ ,  $\tilde{A}^+$  и  $I - \tilde{P}$ , откуда следует, что  $y_3$  ортогонален к  $y_1$  и  $y_2$ . Поэтому

$$\begin{aligned} \|Gx\|^2 &= \|y_1 + y_2\|^2 + \|y_3\|^2 \leq \alpha^2[(\|x_1\| + \|x_2\|)^2 + \|x_1\|^2] = \\ &= \alpha^2[(\cos \varphi + \sin \varphi)^2 + \cos^2 \varphi] = \alpha^2 \left( 1 + \sin 2\varphi + \frac{1 + \cos 2\varphi}{2} \right) = \\ &= \alpha^2 \frac{3 + 2\sin 2\varphi + \cos 2\varphi}{2} \leq \frac{\alpha^2(3 + 5^{1/2})}{2}. \end{aligned}$$

Следовательно,  $\|Gx\| \leq \alpha(1 + 5^{1/2})/2 \approx 1,618\alpha$ .

Тогда

$$\|G\| = \max\{\|Gx\| : \|x\| = 1\} \leq \alpha(1 + 5^{1/2})/2,$$

что доказывает (8.29).

В случае (8.30) либо  $\text{rank } \tilde{A} = n < m$ , и тогда  $\tilde{P} = I_n$ , а  $G_3 = 0$ , либо  $\text{rank } A = m < n$ , и тогда  $Q = I_m$  и  $G_2 = 0$ . Таким образом, либо  $y_2$ , либо  $y_3$  в (8.32) обращаются в нуль, и  $\|Gx\|^2 \leq 2\alpha^2$ . Это устанавливает оценку (8.30) и завершает доказательство теоремы 8.24.

Формулы (8.1)–(8.9) и теоремы этой главы можно использовать для доказательства того, что при соответствующих предположениях о ранге  $A$  элементы матрицы  $A^+$  являются дифференцируемыми функциями элементов  $A$ .

Примеры некоторых конкретных теорем и формул дифференцирования приведены в упражнениях 8.33 и 9.22–9.24. Заметим, что формулы из этих упражнений переносятся на случай, когда  $t$  —  $k$ -мерная переменная с компонентами  $t_1, \dots, t_k$ . Нужно просто заменить  $d/dt$  в этих формулах на  $\partial/\partial t_i$ ,  $i = 1, \dots, k$ .

Дифференцирование псевдообратной матрицы было использовано в [54, 141] для алгоритмов условной минимизации. В [74, 112] дифференцирование псевдообратной матрицы применено к таким нелинейным задачам теории наименьших квадратов, в которые часть параметров входит линейно.

Упражнение 8.33 [90, 137]. Пусть  $A$  —  $m \times n$ -матрица ( $m > n$ ), элементы которой суть дифференцируемые функции действительной переменной  $t$ . Предположим, что при  $t = 0$   $\text{rank } A = n$ . Показать, что: 1) существует действительная окрестность нуля, в которой  $A^+$  является дифференцируемой функцией от  $t$ ; 2) производная  $A^+$  выражается формулой

$$\frac{dA^+}{dt} = -A^+ \frac{dA}{dt} A^+ + A^+ A^+ T \left( \frac{dA}{dt} \right)^T (I - AA^+).$$

## ГЛАВА 9

### ОЦЕНКИ ВОЗМУЩЕНИЙ ДЛЯ РЕШЕНИЙ ЗАДАЧИ НК

Здесь теоремы предыдущей главы будут применены к исследованию воздействия возмущений в  $A$  и  $b$  на решение минимальной длины\*)  $x$  задачи  $Ax \cong b$ . Мы будем по-прежнему пользоваться определениями (8.1)–(8.4). Теорема 9.7 выводится без каких-либо предположений о соотношении размеров  $m$ ,  $n$  и числа  $k = \text{rank } A$ . Вслед за доказательством теоремы 9.7 полученные оценки конкретизируются для трех случаев:  $m = n = k$ ,  $m > n = k$  и  $n > m = k$ .

Удобно формулировать результаты в терминах относительных возмущений

$$\alpha = \frac{\|E\|}{\|A\|}, \quad (9.1)$$

$$\beta = \frac{\|db\|}{\|b\|} \quad (9.2)$$

и величин

$$\gamma = \frac{\|b\|}{\|A\| \|x\|} \leq \frac{\|b\|}{\|Ax\|}, \quad (9.3)$$

$$\rho = \frac{\|r\|}{\|A\| \|x\|} \leq \frac{\|r\|}{\|Ax\|} \quad (r = b - Ax), \quad (9.4)$$

$$\kappa = \|A\| \|A^+\|, \quad (9.5)$$

$$\hat{\kappa} = \frac{\kappa}{1 - \kappa\alpha} \equiv \frac{\|A\| \|A^+\|}{1 - \|A\| \|A^+\|}. \quad (9.6)$$

Определения (9.1)–(9.6) применимы, разумеется, лишь когда соответствующие знаменатели отличны от нуля. Величина  $\kappa$  в (9.5) называется числом обусловленности матрицы  $A$ .

**Теорема 9.7.** Пусть  $x$  — нормальное псевдорешение задачи наименьших квадратов  $Ax \cong b$ ;  $r = b - Ax$  — соответствующий вектор невязки.

\*) В советской литературе приняты термины *нормальное решение* или *нормальное псевдорешение* (смотря по тому, совместна или нет задача  $Ax = b$ ), которые и будут в дальнейшем использоваться. (Примеч. пер.)



Предположим, что  $\|E\| \|A^*\| < 1$  и  $\text{rank } \tilde{A} \leq \text{rank } A$ , и пусть  $x + dx$  — нормальное псевдорешение задачи наименьших квадратов

$$\tilde{A}(x + dx) \equiv (A + E)(x + dx) \equiv b + db.$$

Тогда

$$\text{rank } \tilde{A} = \text{rank } A, \quad (9.8)$$

$$\|dx\| \leq \|A^*\| \left( \frac{\|E\| \|x\|}{1 - \|E\| \|A^*\|} + \frac{\|db\|}{1 - \|E\| \|A^*\|} + \frac{\|E\| \|A^*\| \|r\|}{1 - \|E\| \|A^*\|} + \|E\| \|x\| \right); \quad (9.9)$$

$$\frac{\|dx\|}{\|x\|} \leq \hat{\kappa} \alpha + \hat{\kappa} \gamma \beta + \kappa \hat{\kappa} \rho \alpha + \kappa \alpha \leq \hat{\kappa} [(2 + \kappa \rho) \alpha + \gamma \beta]. \quad (9.10)$$

Доказательство. Равенство (9.8) установлено в теореме 8.15.

Векторы  $x$  и  $x + dx$  выражаются формулами  $x = A^+ b$  и  $x + dx = \tilde{A}^+ (b + db)$ . Поэтому

$$\begin{aligned} dx &= \tilde{A}^+ (b + db) - A^+ b = \\ &= (\tilde{A}^+ - A^+) b + \tilde{A}^+ db = Gb + \tilde{A}^+ db. \end{aligned}$$

Заметим, что  $r = (I - Q)r = (I - Q)b$ , это вместе с (8.8) дает  $G_2 b = G_2 r$ . Из (8.6) — (8.9) следует

$$dx = -\tilde{A}^+ E x + G_2 r + (I - \tilde{P}) E^T A^T x + \tilde{A}^+ db. \quad (9.11)$$

Взяв оценку для  $\|\tilde{A}^+\|$  из теоремы 8.15 и оценку (8.26) для  $\|G_2\|$ , получим неравенство (9.9). Деля это неравенство на  $\|x\|$  и используя определения (9.1) — (9.6), получим неравенство (9.10). Теорема 9.7 доказана.

Заметим, что при  $n = k \equiv \text{rank } A$  матрица  $G_3$  в (8.9) нулевая, поэтому четвертый член в правой части неравенств (9.9) и (9.10) равен нулю. Аналогично при  $m = k \equiv \text{rank } A$  матрица  $G_2$  в (8.8) и, следовательно, третий член в правой части неравенств (9.9) и (9.10) нулевые.

Далее, если  $n = k$  либо  $m = k$ , то  $\text{rank } \tilde{A}$ , очевидно, не может превосходить  $\text{rank } A$ . Таким образом, в этих случаях предположение  $\text{rank } \tilde{A} < \text{rank } A$ , использовавшееся в теореме 9.7, выполняется автоматически. Сделанные замечания доказывают следующие три теоремы.

**Теорема 9.12.** Предположим, что  $m > n = k \equiv \text{rank } A$  и  $\|E\| \|A^*\| < 1$ . Тогда

$$\text{rank } \tilde{A} = \text{rank } A,$$

$$\|dx\| \leq \frac{\|A^*\| (\|E\| (\|x\| + \|A^*\| \|r\|) + \|db\|)}{1 - \|E\| \|A^*\|}, \quad (9.13)$$

$$\frac{\|dx\|}{\|x\|} \leq \hat{\kappa} [(1 + \kappa \rho) \alpha + \gamma \beta]. \quad (9.14)$$

**Теорема 9.15.** Предположим, что  $m = n = k \equiv \text{rank } A$  и  $\|E\| \|A^*\| < 1$ . Тогда

$$\text{rank } \tilde{A} = \text{rank } A,$$

$$\|dx\| \leq \frac{\|A^*\| (\|E\| \|x\| + \|db\|)}{1 - \|E\| \|A^*\|}, \quad (9.16)$$

$$\frac{\|dx\|}{\|x\|} \leq \hat{\kappa} (\alpha + \gamma \beta) \leq \hat{\kappa} (\alpha + \beta). \quad (9.17)$$

**Теорема 9.18.** Предположим, что  $n > m = k \equiv \text{rank } A$  и  $\|E\| \|A^*\| < 1$ . Тогда

$$\text{rank } \tilde{A} = \text{rank } A,$$

$$\|dx\| \leq \|A^*\| \left( \frac{\|E\| \|x\| + \|db\|}{1 - \|E\| \|A^*\|} + \|E\| \|x\| \right), \quad (9.19)$$

$$\frac{\|dx\|}{\|x\|} \leq \tilde{\kappa} (\alpha + \gamma \beta) + \kappa \alpha \leq \hat{\kappa} (2\alpha + \gamma \beta) \leq \hat{\kappa} (2\alpha + \beta). \quad (9.20)$$

Применяя доказательство, аналогичное выводу (8.30), можно показать, что

$$\frac{\|dx\|}{\|x\|} \leq \hat{\kappa} (2^{1/2} \alpha + \beta). \quad (9.21)$$

В приводимых ниже упражнениях даны примеры формул дифференцирования, выводимых из результатов этой главы. Обобщения и приложения этих формул указаны в конце гл. 8.

#### Упражнения

9.22. Пусть  $A$  —  $m \times n$ -матрица ( $m > n$ ), элементы которой суть дифференцируемые функции действительной переменной  $t$ . Пусть  $x$  — вектор-функция от  $t$ , определяемая условием  $Ax \equiv b$  для всех  $t$  в той окрестности  $U$ , где  $A$  дифференцируема и  $\text{rank } A = n$ . Показать, что для  $t \in U$  существует  $dx/dt$ , которая является решением задачи наименьших квадратов

$$A \left( \frac{dx}{dt} \right) \equiv - \left( \frac{dA}{dt} \right) x + A^T \left( \frac{dA}{dt} \right)^T r,$$

где  $r = b - Ax$ .

9.23. Показать далее, что  $dx/dt$  есть решение квадратной невырожденной системы

$$A^T A \left( \frac{dx}{dt} \right) = -A^T \left( \frac{dA}{dt} \right) x + \left( \frac{dA}{dt} \right)^T r.$$

9.24. Пусть  $A = Q^T R$ , где  $Q$  —  $n \times m$ -матрица с ортонормальными строками, а  $R$  — невырожденная  $n \times n$ -матрица (разложение  $A$  получено, например, посредством преобразований Хаусхолдера). Показать, что  $dx/dt$  удовлетворяет уравнению

$$R \left( \frac{dx}{dt} \right) = -Q \left( \frac{dA}{dt} \right) x + (R^{-1})^T \left( \frac{dA}{dt} \right)^T r.$$

# ВЫЧИСЛЕНИЯ, ИСПОЛЬЗУЮЩИЕ ЭЛЕМЕНТАРНЫЕ ОРТОГОНАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ

Мы переходим теперь к описанию численных алгоритмов для ортогонального разложения  $m \times n$ -матрицы  $A$  (указанного в теореме 3.19) и решения задачи НК.

Так как преобразования Хаусхолдера и Гивенса имеют очень много приложений, мы сочли, что будет удобно и полезно с точки зрения унификации рассмотреть каждое из этих преобразований как самостоятельный объект. Оба преобразования будут сейчас разобраны в деталях. Эти модули в дальнейшем используются в описании других, более сложных вычислительных процедур.

Вычисление преобразования Хаусхолдера можно разбить на два этапа: 1) построение преобразования; 2) его применение к другим векторам. Ортогональное  $m \times m$ -преобразование Хаусхолдера можно представить в виде

$$Q = I_m + b^{-1}uu^T, \quad (10.1)$$

где  $u$  —  $m$ -вектор такой, что  $\|u\| \neq 0$ , и  $b = -\|u\|^2/2$ .

В гл. 3 было показано, что для данного вектора  $u$  можно определить вектор  $u$  таким образом, чтобы выполнялось (3.2). На практике встречаются ситуации, когда  $Q$  нужно определить из условия

$$Qv = \begin{bmatrix} v_1 \\ \dots \\ v_{p-1} \\ -\sigma \left( v_p^2 + \sum_{i=1}^m v_i^2 \right)^{1/2} \\ v_{p+1} \\ \dots \\ v_{l-1} \\ 0 \\ \dots \\ 0 \end{bmatrix} \equiv y. \quad (10.2)$$

Построить такую матрицу  $Q$  можно было бы как произведение перестановки строк, обычного преобразования Хаусхолдера, описанного в лемме 3.1, и обратной перестановки строк. Мы находим, однако, более удобным рассматривать равенство (10.2) как определяющее цель нашего основного вычислительного модуля. Действие матрицы  $Q$  при преобразовании  $u$  в  $y$  можно описать с помощью трех неотрицательных целых параметров  $p, l, m$  следующим образом:

1) если  $p > 1$ , то компоненты с номерами  $1, \dots, p-1$  не должны меняться;

2) компонента с номером  $p$  может измениться. Она называется главным элементом;

3) если  $p < l-1$ , то компоненты с номерами  $p+1, \dots, l-1$  не должны меняться;

4) если  $l \leq m$ , то компоненты с номерами  $l, \dots, m$  должны быть аннулированы.

Заметим, что должны выполняться соотношения

$$1 \leq p \leq m, \quad (10.3)$$

$$p < l. \quad (10.4)$$

Шаги численного процесса, который приводит к ортогональной  $m \times m$ -матрице  $Q$ , отвечающей целым параметрам  $p, l, m$ , можно свести в следующую схему:

$$s = -\sigma \left( v_p^2 + \sum_{i=1}^m v_i^2 \right)^{1/2}, \quad (10.5)$$

$$\sigma = \begin{cases} +1, & v_p \geq 0, \\ -1, & v_p < 0, \end{cases} \quad (10.6)$$

$$u_i = 0, \quad i = 1, \dots, p-1, \quad (10.7)$$

$$u_p = v_p - s, \quad (10.8)$$

$$u_i = 0, \quad i = p+1, \dots, l-1, \quad (10.9)$$

$$u_i = v_i, \quad i = l, \dots, m, \quad (10.10)$$

$$b = su_p, \quad (10.11)$$

$$Q = \begin{cases} I_m + b^{-1}uu^T, & b \neq 0, \\ I_m, & b = 0. \end{cases} \quad (10.12)$$

Тот факт, что матрица  $Q$ , определенная в (10.11), имеет нужные свойства, устанавливается следующими тремя леммами.

**Л е м м а 10.12.** Определяемые формулами (10.5) — (10.10)  $m$ -вектор  $u$  и скаляр  $b$  удовлетворяют равенству

$$b = -\|u\|^2/2. \quad (10.13)$$

**Доказательство.** Имеем

$$\begin{aligned} \|u\|^2 &= (v_p - s)^2 + \sum_{j=1}^m v_j^2 = \\ &= v_p^2 - 2v_p s + s^2 + \sum_{j=1}^m v_j^2 = -2s(v_p - s) = -2su_p = -2b. \end{aligned}$$

Лемма доказана.

**Л е м м а 10.14.** Матрица  $Q$ , определяемая формулой (10.11), ортогональная.

**Доказательство.** Проверка условия  $Q^T Q = I_m$  проводится "в лоб" с использованием (10.11) и (10.13).

**Лемма 10.15.** Пусть  $y = Qv$ . Тогда

$$y_i = v_i, \quad i = 1, \dots, p-1, \quad (10.16)$$

$$y_p = s, \quad (10.17)$$

$$y_i = v_i, \quad i = p+1, \dots, l-1, \quad (10.18)$$

$$y_i = 0, \quad i = l, \dots, m. \quad (10.19)$$

**Доказательство.** Если  $v = 0$ , то лемма очевидным образом верна. При  $v \neq 0$  легко проверяется с использованием (10.6) – (10.9), что  $u^T v = -b$ . Но тогда вектор

$$y = Qv = v - u$$

удовлетворяет равенствам (10.16) – (10.19).

В реальных вычислениях ненулевые компоненты векторов  $u$  и  $y$  обычно размещают в массиве памяти, который прежде занимал вектор  $u$ . Исключением является  $p$ -я компонента массива; приходится выбирать, поместить ли в нее  $u_p$  или  $y_p$ . Мы остановимся на втором варианте, а для хранения  $u_p$  заведем дополнительную ячейку. Величина  $b$  может быть вычислена в соответствии с (10.10) всякий раз, как она нужна.

После того как преобразование построено, обычно требуется применить его к некоторому набору  $m$ -векторов  $c_1, \dots, c_\nu$ , т.е. нужно вычислить векторы  $\tilde{c}_j = Qc_j$ ,  $j = 1, \dots, \nu$ . Используя определение  $Q$  в (10.11), это вычисление можно выполнять так:

$$t_j = b^{-1}(u^T c_j), \quad (10.20)$$

$$\tilde{c}_j = c_j + t_j u, \quad j = 1, \dots, \nu. \quad (10.21)$$

Все вышеизложенное будет теперь переформулировано в алгоритмической форме, пригодной для реализации машинной программой. Мы опишем алгоритм Н1 ( $p, l, m, v, h, C, \nu$ ) для построения и (по желанию пользователя) применения преобразования Хаусхолдера и алгоритм Н2 ( $p, l, m, v, h, C, \nu$ ) для применения (по желанию пользователя) построенного ранее преобразования Хаусхолдера.

Входной информацией алгоритма Н1 являются целые числа  $p, l, m$  и  $\nu$ ,  $m$ -вектор  $v$  и при  $\nu > 0$  массив  $C$ , содержащий  $m$ -векторы  $c_j$ ,  $j = 1, \dots, \nu$ .

Массив  $C$  может иметь либо размеры  $m \times \nu$ , и в этом случае векторы  $c_j$  являются его столбцами, либо размеры  $\nu \times m$ , и тогда  $c_j$  будут его строками. В описании алгоритмов Н1 и Н2 мы не будем различать эти два возможных способа хранения. Однако при ссылках на эти алгоритмы в последующем тексте книги хранение по столбцам будет считаться стандартным методом хранения. В тех же случаях, когда векторы  $c_j$ , к которым применяется алгоритм Н1 (или Н2), хранятся по строкам  $C$ , это будет специально оговариваться.

Алгоритм Н1 вычисляет вектор  $u$ , число  $b$ , вектор  $y = Qv$  и при  $\nu > 0$  векторы  $\tilde{c}_j = Qc_j$ ,  $j = 1, \dots, \nu$ . Выходными данными алгоритма Н1 являются:

ся:  $p$ -я компонента  $u$ , хранящаяся в ячейке  $h$ ; компоненты  $u$  с номерами  $1, \dots, m$ , хранимые в одноименных позициях массива  $v$ ; компоненты  $y$  с номерами  $1, \dots, l-1$ , хранимые в одноименных позициях массива  $v$ , и при  $\nu > 0$  векторы  $\tilde{c}_j$ ,  $j = 1, \dots, \nu$ , хранимые в массиве  $C$ .

В алгоритме Н2 величины  $p, l$  и  $m$  сохраняют тот же смысл, что и в алгоритме Н1. Вектор  $v$  и величина  $h$  должны иметь значения, вычисленные в результате предварительного выполнения алгоритма Н1. Эти данные определяют матрицу преобразования  $Q$ . Если  $\nu > 0$ , то на входе в алгоритм массив  $C$  должен содержать  $m$ -векторы  $c_j$ ,  $j = 1, \dots, \nu$ . Алгоритм Н2 заменит их векторами  $\tilde{c}_j = Qc_j$ ,  $j = 1, \dots, \nu$ .

**Алгоритмы 10.22.** Н1 ( $p, l, m, v, h, C, \nu$ ) (выполняются шаги 1–11), Н2 ( $p, l, m, v, h, C, \nu$ ) (выполняются шаги 5–11):

$$1. \text{ Положить } s := (v_p^2 + \sum_{i=1}^m v_i^2)^{1/2}.$$

$$2. \text{ Если } v_p > 0, \text{ положить } s := -s.$$

$$3. \text{ Положить } h := v_p - s, \quad v_p := s.$$

4. **Комментарий.** Построение преобразования закончено. На шаге 5 начинается применение преобразования к векторам  $c_j$ .

$$5. \text{ Положить } b := v_p h.$$

$$6. \text{ Если } b = 0 \text{ или } \nu = 0, \text{ перейти к шагу 11.}$$

$$7. \text{ Для } j := 1, \dots, \nu \text{ выполнить шаги 8–10.}$$

$$8. \text{ Положить } s := (c_{pj}h + \sum_{i=1}^m c_{ij}v_i)/b.$$

$$9. \text{ Положить } c_{pj} := c_{pj} + sh.$$

$$10. \text{ Для } i = 1, \dots, m \text{ положить } c_{ij} = c_{ij} + sv_i.$$

11. **Комментарий.**

а) Алгоритм Н1 (или Н2) закончен.

б) На шаге 1 вычисление квадратного корня из суммы квадратов можно защитить от возможности получения машинного нуля, если использовать тождество  $(w_1^2 + \dots + w_m^2)^{1/2} = t [(w_1/t)^2 + \dots + (w_m/t)^2]^{1/2}$ , где  $t = \max\{|w_i|, i = 1, \dots, m\}$ .

Другое универсальное элементарное ортогональное преобразование, которое мы обсудим, – это вращение Гивенса. Формулы (3.5) – (3.9) описывают его построение. Чтобы избежать неоправданных машинных нулей или переполнений, можно вычислять квадратный корень  $r = (x^2 + y^2)^{1/2}$  в формулах для  $c$  и  $s$  в соответствии со следующим предписанием:

$$t = \max(|x|, |y|), \quad u = \min(|x|, |y|),$$

$$r = \begin{cases} t[1 + (u/t)^2]^{1/2}, & t \neq 0, \\ 0, & t = 0. \end{cases} \quad (10.23)$$

Если арифметика вашей машины отличается от нормализованной двоичной арифметики, то следует принять дополнительные меры предосторожности, чтобы не потерять точность в вычислении  $c$  и  $s$ . Например, в [39] ука-

зано такое преобразование формулы (10.23) для работы с нормализованной шестнадцатеричной арифметикой:

$$r = 2t \left[ \frac{1}{4} + \left( \frac{u}{2t} \right)^2 \right]^{1/2} \quad (10.24)$$

Сейчас будут описаны алгоритмы  $G1(v_1, v_2, c, s, r)$  и  $G2(c, s, z_1, z_2)$ , предназначенные соответственно для построения и применения вращения Гивенса. При этом будет использоваться формула (10.23), отвечающая арифметике с основанием 2. Входные данные алгоритма  $G1$  — это компоненты  $v_1$  и  $v_2$  2-вектора  $v$ . На выходе будут получены скаляры  $c$  и  $s$ , определяющие в соответствии с (3.5) матрицу  $G$ , а также квадратный корень из суммы квадратов  $v_1$  и  $v_2$ , хранимый в ячейке  $r$ . Ячейку  $r$  можно отождествить с ячейками, хранящими  $v_1$  либо  $v_2$ ; первый вариант обычно самый удобный. Входная информация алгоритма  $G2$  состоит из скаляров  $c$  и  $s$ , определяющих матрицу  $G$  по формуле (3.5), и компонент  $z_1$  и  $z_2$  2-вектора  $z$ . На выходе будут получены компоненты  $d_1$  и  $d_2$  вектора  $d = Gz$ ; они помещены в ячейки  $z_1$  и  $z_2$ .

Алгоритм 10.25.  $G1(v_1, v_2, c, s, r)$ :

1. Если  $|v_1| \leq |v_2|$ , перейти к шагу 8.
2. Положить  $w := v_2/v_1$ .
3. Положить  $q := (1 + w^2)^{1/2}$ .
4. Положить  $c := 1/q$ .
5. Если  $v_1 < 0$ , положить  $c := -c$ .
6. Положить  $s := wc$ .
7. Положить  $r := |v_1|q$  и перейти к шагу 16.
8. Если  $v_2 \neq 0$ , перейти к шагу 10.
9. Положить  $c := 1, s := 0, r := 0$  и перейти к шагу 16.
10. Положить  $w := v_1/v_2$ .
11. Положить  $q := (1 + w^2)^{1/2}$ .
12. Положить  $s := 1/q$ .
13. Если  $v_2 < 0$ , положить  $s := -s$ .
14. Положить  $c := ws$ .
15. Положить  $r := |v_2|q$ .
16. Комментарий. Преобразование построено.

Алгоритм 10.26.  $G2(c, s, z_1, z_2)$ :

1. Положить  $w := z_1c + z_2s$ .
2. Положить  $z_2 := -z_1s + z_2c$ .
3. Положить  $z_1 := w$ .

4. Комментарий. Применение преобразования закончено.

При реализации преобразований Хаусхолдера или Гивенса возможно большое разнообразие алгоритмических и программистских деталей. Это связано с тем, что во главу угла могут быть поставлены различные соображения: скорость выполнения, точность, защищенность от машинных нулей или переполнений, сокращение требований к памяти, уменьшение сложности программы, модульность программы, использование разреженности ненулевых элементов, язык программирования, транспортабельность и т.д. Мы приведем сейчас два примера других реализаций.

Обсуждая преобразование Хаусхолдера, мы исходили из представления (10.1). Однако матрицу Хаусхолдера можно записать и так:

$$Q = I + gh^T = I + \begin{bmatrix} g_1 \\ \dots \\ g_m \end{bmatrix} \cdot [1 \quad h_2 \quad \dots \quad h_m]. \quad (10.27)$$

Формулы (10.1) и (10.27) связаны подстановками  $g = b^{-1}u, u = s^{-1}u$ ,  $h = u_1^{-1}u$ .

Формула (10.27) представляет интерес главным образом для малых  $m$ . В этом случае необходимость хранить два вектора  $g$  и  $h$  вместо одного вектора  $u$  не так обременительна; зато экономия двух умножений всякий раз, как вычисляется произведение  $\tilde{c} = Qc$ , приобретает некоторое значение. Эта форма использована при  $m = 3$  в алгоритме Мартина, Питерса и Уилкинсона [8] и при  $m = 2, 3$  в алгоритме Моулера и Стьюарта [127].

Для  $m = 2$  вычисление  $\tilde{c} = Qc$  посредством (10.1) (см. формулы (10.20), (10.21)) требует пяти умножений (или четырех умножений и одного деления) и трех сложений, в то время как при использовании (10.27) нужны лишь три умножения и три сложения. Таким образом, (10.27) требует меньшего числа операций, чем (10.1). Более того, при применении (10.27) выдерживается конкуренция и с преобразованием Гивенса, которое, будучи реализовано алгоритмом  $G2$  (см. 10.26), требует для вычисления  $\tilde{c} = Qc$  четырех умножений и двух сложений.

Реальное соотношение характеристик машинной программы, основанной на (10.27), и программы обычного преобразования Гивенса [алгоритмы  $G1$  (см. 10.25) и  $G2$  (см. 10.26)], скорей всего, зависит от деталей этих программ. Сошлемся в качестве примера на свой опыт проверки на машине UNIVAC 1108 двух таких программ, использующих арифметику обыкновенной точности (27-битовая мантисса). Эти программы применялись для аннулирования элемента (2, 1) в 100 различных  $2 \times 11$ -матрицах, также заданных с обыкновенной точностью. Для каждого из 1100 преобразованных 2-векторов вычислялась относительная погрешность  $\rho = \|v' - v''\| / \|v'\|$ , где  $v'$  — преобразованный вектор, полученный проверяемой программой при использовании обыкновенной точности, а  $v''$  — преобразованный вектор, вычисленный той же программой в версии с двойной точностью. Для подпрограммы Хаусхолдера, базирующейся на представлении (10.27), среднееквадратичное значение  $\rho$  было равно  $1,72 \times 2^{-27}$ , а максимальное значение  $\rho$  равнялось  $8,70 \times 2^{-27}$ . Для подпрограммы Гивенса соответствующие значения были  $0,88 \times 2^{-27}$  и  $3,17 \times 2^{-27}$ .

В работах [65, 66] были предложены методы сокращения числа операций в преобразовании Гивенса. В этих методах требуется, чтобы для хранения обеих матриц — матрицы  $A$ , на которую воздействует преобразование, и преобразованной матрицы  $\tilde{A} = GA$  — использовалась факторизованная форма.

Сейчас будет описан один из методов Джентльмена. Чтобы облегчить описание, будем считать  $A$   $2 \times n$ -матрицей, в которой нужно аннулировать элемент (2, 1) посредством левого умножения на вращение Гивенса. Таким

образом,

$$GA = \tilde{A}, \quad (10.28)$$

где  $\tilde{a}_{21} = 0$ . Если бы мы строили матрицу Гивенса  $G$  в явном виде, то следовало бы вычислить

$$r = (a_{11}^2 + a_{21}^2)^{1/2} \quad (10.29)$$

и положить

$$G = \begin{cases} \begin{bmatrix} \frac{a_{11}}{r} & \frac{a_{21}}{r} \\ -\frac{a_{21}}{r} & \frac{a_{11}}{r} \end{bmatrix}, & r > 0, \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & r = 0. \end{cases} \quad (10.30)$$

В рассматриваемом методе вместо матрицы  $A$  в памяти хранятся матрицы  $D_{2 \times 2}$  и  $B_{2 \times n}$  такие, что

$$D = \text{Diag}\{d_1, d_2\}, \quad d_i > 0, \quad (10.31)$$

$$A = D^{1/2} B. \quad (10.32)$$

Мы хотим заменить в памяти матрицы  $D$  и  $B$  новыми матрицами  $\tilde{D}$  и  $\tilde{B}$ , причем

$$\tilde{D} = \text{Diag}\{\tilde{d}_1, \tilde{d}_2\}, \quad \tilde{d}_i > 0, \quad (10.33)$$

такова, что матрица  $\tilde{A}$  из (10.28) представима в виде

$$\tilde{A} = \tilde{D}^{1/2} \tilde{B} \quad \text{или} \quad \tilde{A} = -\tilde{D}^{1/2} \tilde{B}. \quad (10.34)$$

Будем различать три случая:

1.  $b_{21} = 0$ .
2.  $0 < d_2 b_{21}^2 < d_1 b_{11}^2$ .
3.  $0 < d_1 b_{11}^2 < d_2 b_{21}^2$ .

Случай 1. Можно положить  $\tilde{D} = D$  и  $\tilde{B} = B$ .

Случай 2. Положим

$$t = \frac{d_2 b_{21}^2}{d_1 b_{11}^2}, \quad (10.35)$$

$$\tilde{d}_1 = \frac{d_1}{1+t}, \quad (10.36)$$

$$\tilde{d}_2 = \frac{d_2}{1+t}, \quad (10.37)$$

$$H = \begin{bmatrix} 1 & \frac{d_2 b_{21}}{d_1 b_{11}} \\ -\frac{b_{21}}{b_{11}} & 1 \end{bmatrix} \quad (10.38)$$

$$\tilde{B} = HB. \quad (10.39)$$

Заметим, что  $\tilde{b}_{11} = b_{11}(1+t)$  и  $\tilde{b}_{21} = 0$ .

Случай 3. Положим

$$t = \frac{d_1 b_{11}^2}{d_2 b_{21}^2}, \quad (10.40)$$

$$\tilde{d}_1 = \frac{d_2}{1+t}, \quad (10.41)$$

$$\tilde{d}_2 = \frac{d_1}{1+t}, \quad (10.42)$$

$$H = \begin{bmatrix} \frac{d_1 b_{11}}{d_2 b_{21}} & 1 \\ -1 & \frac{b_{11}}{b_{21}} \end{bmatrix} \quad (10.43)$$

$$\tilde{B} = HB. \quad (10.44)$$

Заметим, что  $\tilde{b}_{11} = b_{21}(1+t)$  и  $\tilde{b}_{21} = 0$ .

Легко проверить, что определяемые этим процессом матрицы  $\tilde{D}$  и  $\tilde{B}$  удовлетворяют условию (10.34). Экономия в арифметике при применении преобразования в этой форме происходит из присутствия двух единиц в матрице  $H$ . Это позволяет перемножить матрицы  $\tilde{B} = HB$  за  $2n$  сложений и  $2n$  умножений по сравнению с  $2n$  сложениями и  $4n$  умножениями при матричном умножении  $\tilde{A} = GA$ . Кроме того, удалось избавиться от операции извлечения квадратного корня, обычно используемой при построении преобразования Гивенса.

Если произвольная  $m \times n$ -матрица  $A$  посредством последовательности вращений Гивенса приводится к треугольному виду, то можно начинать с  $D_1 = I_{m \times m}$ ,  $B_1 = A$  и строить с помощью описанной процедуры последовательность матриц  $\{D_k\}$ ,  $\{B_k\}$ ,  $k = 1, 2, \dots$ . Элементы матриц  $D_k$  обычно уменьшаются с ростом  $k$ , но скорость убывания ограничена, поскольку  $1/2 < (1+t)^{-1} < 1$ . Соответствующий, хотя и несколько более медленный, рост имеет место в элементах матриц  $B_k$ . Причина роста в том, что евклидова норма каждого столбца произведения  $D_k^{1/2} B_k$  инвариантна относительно  $k$ .

Чтобы уменьшить вероятность машинных нулей в элементах  $D$  или переполнений в элементах  $B$ , всякая программа, реализующая этот вариант вращений Гивенса, должна содержать некоторую процедуру, контролирующую величину чисел  $d_i$  и изменяющую масштабирование  $d_i$  и  $i$ -й строки  $B$ , если  $d_i$  становится меньше некоторого заданного допуска. Например, можно положить  $\tau = 2^{-24}$ ,  $\rho = \tau^{-1}$  и  $\beta = \tau^{1/2}$ . Если нужно оперировать с числом  $d_i$ , его можно вначале сравнить с  $\tau$ . Если  $d_i < \tau$ , то  $d_i$  заменяется числом  $\rho d_i$ , а элементы  $b_{ij}$  — значениями  $\beta b_{ij}$ ,  $j = 1, \dots, n$ . При указанном выборе  $\tau$  умножения на  $\rho$  и  $\beta$  будут точными операциями на многих машинах, имеющих арифметику с плавающей запятой и основанием 2, 8 или 16.

## ГЛАВА 11

### ВЫЧИСЛЕНИЕ РЕШЕНИЯ ПЕРЕОПРЕДЕЛЕННОЙ ИЛИ ТОЧНО ОПРЕДЕЛЕННОЙ ЗАДАЧИ ПОЛНОГО РАНГА

Теорема 3.11 показывает, что для  $m \times n$ -матрицы  $A$  существует ортогональная  $m \times m$ -матрица  $Q$  такая, что матрица  $QA = R$  нулевая ниже главной диагонали. В этой главе мы опишем способ вычисления матриц  $Q$  и  $R$  с помощью алгоритмов Н1 и Н2 (см. гл. 10).

Матрица  $Q$  есть произведение преобразований Хаусхолдера

$$Q = Q_n \dots Q_1, \quad (11.1)$$

где каждый сомножитель  $Q_j$  имеет форму

$$Q_j = I_m + b_j^{-1} u^{(j)} u^{(j)T}, \quad j = 1, \dots, n. \quad (11.2)$$

Вместо того чтобы хранить величины  $b_j$ , участвующие в (11.2), мы будем перевычислять их в случае необходимости по формуле (10.10).

Вводя индексы, запишем

$$b_j = s_j u_j^{(j)}, \quad j = 1, \dots, n. \quad (11.3)$$

Каждая величина  $s_j$  — это  $j$ -й диагональный элемент матрицы  $R$  и будет храниться в качестве такового. Величины  $u_j^{(j)}$  хранятся во вспомогательном массиве ячеек с именами  $h_j$ ,  $j = 1, \dots, n$ .

Наш алгоритм вычислит разложение теоремы 3.11. Он будет именоваться НФТ ( $m, n, A, h$ ). На вход НФТ подаются целые числа  $m, n$  и  $m \times n$ -матрица  $A$ . Выходная информация состоит из ненулевой части верхней треугольной матрицы  $R$ , хранимой в верхней треугольной части массива с именем  $A$ , скаляров  $u_j^{(j)}$ , хранимых в массиве с именем  $h$ ; прочие ненулевые компоненты векторов  $u^{(j)}$  хранятся по столбцам поддиагональной части массива  $A$ .

Алгоритм 11.4. НФТ ( $m, n, A, h$ ):

1. Для  $j := 1, \dots, n$  выполнить алгоритм Н1 ( $j, j+1, m, a_{1j}, h_j, a_{1,j+1}, n-j$ ) (см. 10.22).

2. Комментарий. Приведение к треугольному виду закончено.

На шаге 1 описанного алгоритма мы применяем договоренность (согласованную с практикой использования фортрана), согласно которой на  $j$ -й столбец  $A$  можно сослаться указанием имени его первой компоненты, а на подматрицу, образованную столбцами  $j+1, \dots, n$  матрицы  $A$ , можно сослаться через имя первого элемента столбца  $j+1$ .

В случаях 1а и 2а на рис. 1.1 верхняя треугольная  $n \times n$ -матрица  $R_{11}$  (см. (3.21)) невырождена. Поэтому в этих двух случаях решение  $\hat{x}$  задачи НК можно получить, вычисляя вектор

$$g = Qb, \quad (11.5)$$

представляя  $g$  в виде

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix} \quad (11.6)$$

и решая систему

$$R_{11}x = g_1 \quad (11.7)$$

относительно вектора  $\hat{x}$ .

Исходя из (2.9) и (2.10), вектор невязки  $r = b - A\hat{x}$  и его норму можно вычислить так:

$$r = Q_1 \dots Q_n \begin{bmatrix} 0 \\ g_2 \end{bmatrix}, \quad (11.8)$$

$$\rho \equiv \|r\| = \|g_2\|. \quad (11.9)$$

Формулы (11.8) и (11.9) устраняют необходимость в хранении или восстановлении исходных данных задачи  $[A:b]$  для вычисления невязок.

Приводимый ниже алгоритм HS1 выполняет вычисления согласно (11.5)–(11.7). Входная информация этого алгоритма состоит из целых чисел  $m$  и  $n$ , массивов с именами  $A$  и  $h$  в том виде, как они получены алгоритмом НФТ (см. 11.4), и массива  $b$ , хранящего правую часть задачи НК —  $m$ -вектор  $b$ .

На выходе алгоритма будут получены  $n$ -вектор  $x$ , замещающий первые  $n$  компонент массива  $b$ , и (при  $m > n$ )  $m-n$ -вектор  $g_2$ , замещающий компоненты массива  $b$  с номерами  $n+1, \dots, m$ .

Алгоритм 11.10. HS1 ( $m, n, A, h, b$ ):

1. Для  $j := 1, \dots, n$  выполнить алгоритм Н2 ( $j, j+1, m, a_{1j}, h_j, b, 1$ ).

2. Комментарий. На шагах 3–5 будет вычислено решение треугольной системы  $R_{11}x = g_1$  из (11.7).

3. Положить  $b_n := b_n/a_{nn}$ .

4. Если  $n \leq 1$ , перейти к шагу 6.

5. Для  $i := n-1, n-2, \dots, 1$  положить  $b_i := \frac{1}{a_{ii}} \left( b_i - \sum_{j=i+1}^n a_{ij} b_j \right)$ .

6. Комментарий. Для случаев 1а и 2а вычислено решение задачи НК.

Алгоритм легко приспособляется к случаю, когда  $b$  является  $m \times l$ -матрицей. Достаточно в описании шага 1 заменить значение последнего параметра

ра алгоритма Н2 на  $l$  и изменить шаги 3 и 5 так, чтобы  $b$  обрабатывался как массив размера  $m \times l$ .

Чтобы вычислить норму  $\rho$  невязки (см. (11.9)), можно добавить к алгоритму HS1 шаг

$$7. \rho = \left( \sum_{i=n+1}^m b_i^2 \right)^{1/2}.$$

Чтобы вычислить вектор невязки (см. (11.8)), можно добавить к алгоритму HS1 шаг

8. Для  $i = 1, \dots, n$  положить  $b_i = 0$ .

9. Для  $j = n, n-1, \dots, 1$  выполнить алгоритм Н2( $j, j+1, m, a_{1j}, h_j, b, 1$ ).

Заметим, что если нужно сохранить вектор решения  $x$ , то перед выполнением шагов 8 и 9 его следует переместить из массива  $b$  в какой-либо другой массив. После окончания шага 9 в массиве  $b$  будет находиться вектор невязки  $r = b - Ax$ .

Алгоритм HS1 основан на предположении, что матрица  $A$  задачи НК имеет ранг  $n$ . В алгоритме не делается проверки этого предположения.

На практике важно знать, могут ли изменения матрицы  $A$  порядка неопределенности в исходных данных привести к тому, что ее ранг станет меньше  $n$ . Один из численных подходов к этой задаче, связанный с использованием в качестве первого шага перестановок столбцов, будет обсуждаться в гл. 14. Другой метод, дающий более детальную информацию о матрице, требует вычисления сингулярного разложения. Он будет рассмотрен в гл. 18.

Часто при решении задачи наименьших квадратов интерес представляет также вычисление матрицы ковариации для вектора решения. Этому вопросу посвящена гл. 12.

#### Упражнения

11.11. Описать алгоритм приведения матрицы к треугольному виду, основанный на использовании преобразований Гивенса. Подсчитать отдельно число сложений и число умножений. Как изменятся эти числа, если использовать преобразования Гивенса в быстрой форме (два умножения, два сложения)?

11.12. Определить число умножений, необходимых для решения задачи НК посредством алгоритмов НФТ и HS1. Сравнить это число с тем, что получено в упражнении 11.11 для преобразований Гивенса.

## ГЛАВА 12

### ВЫЧИСЛЕНИЕ КОВАРИАЦИОННОЙ МАТРИЦЫ РЕШЕНИЯ

Симметричная, положительно определенная матрица

$$C = (A^T A)^{-1}, \text{ rank } A = n, \quad (12.1)$$

или ее скалярное кратное (скажем,  $\sigma^2 C$ ), имеет при соответствующих предположениях статистическую интерпретацию в качестве оценки ковариационной матрицы для решения задачи НК (см., например, [144]). Мы будем называть  $C$  *нешкалированной ковариационной матрицей*.

В этой главе будут представлены некоторые алгоритмы вычисления  $C$ . Мы выделим также некоторые часто встречающиеся ситуации, когда можно избежать явного вычисления  $C$ , пересматривая ее роль в последующем анализе.

Мы не станем обсуждать причину появления или интерпретацию скалярного множителя  $\sigma^2$ . Отметим только, что обычно  $\sigma^2$  вычисляют по формуле

$$\sigma^2 = \frac{\|A\hat{x} - b\|}{m - n}, \quad (12.2)$$

где  $\hat{x}$  — решение задачи  $Ax \cong b$ , а  $m$  и  $n$  — соответственно число строк и число столбцов матрицы  $A$ .

Мы обсудим алгоритмы вычисления  $C$ , использующие те или иные разложения матрицы  $A$ . Эти разложения вычисляются в ходе трех различных методов решения задачи НК, детально изучаемых в данной книге, и выглядят следующим образом:

$$QA = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad R_{n \times n} \text{ — верхняя треугольная (см. гл. 11);} \quad (12.3)$$

$$\tilde{Q}AP = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix}, \quad \tilde{R}_{n \times n} \text{ — верхняя треугольная (см. гл. 14);} \quad (12.4)$$

$$U^T A V = \begin{bmatrix} S \\ 0 \end{bmatrix}, \quad S_{n \times n} \text{ — диагональная (см. гл. 18).} \quad (12.5)$$

Разложения (12.3)–(12.5) вычисляются соответственно алгоритмом НФТ (см. 11.4), алгоритмом НФТИ (см. 14.9) и алгоритмом сингулярного разложения гл. 18. Матрица  $Q$  в (12.3) ортогональная порядка  $m$ . Обе матрицы  $\tilde{Q}$  и  $P$  в (12.4) ортогональные, причем  $P$  — матрица перестановки. Матрицы  $U$  и  $V$  в (12.5) имеют порядок  $m$  и  $n$  соответственно, и обе они ортогональны.

Легко видеть, что соответственно этим трем случаям

$$A^T A = R^T R, \quad (12.6)$$

$$A^T A = P^T \tilde{R}^T \tilde{R} P, \quad (12.7)$$

$$A^T A = V S^2 V^T, \quad (12.8)$$

и если  $\text{rank } A = n$ , то обращение этих уравнений дает

$$C = (A^T A)^{-1} = R^{-1} (R^{-1})^T, \quad (12.9)$$

$$C = (A^T A)^{-1} = P \tilde{R}^{-1} (\tilde{R}^{-1})^T P^T, \quad (12.10)$$

$$C = (A^T A)^{-1} = V S^{-2} V^T. \quad (12.11)$$

Рассмотрим вначале детали вычислений в соответствии с формулами (12.9) и (12.10). Здесь три основных шага.

1. Обращение верхней треугольной матрицы  $R$  (или  $\tilde{R}$ ) на месте, занимаемом ею в памяти.

2. Формирование верхней треугольной части симметричной матрицы  $R^{-1} (R^{-1})^T$  или  $\tilde{R}^{-1} (\tilde{R}^{-1})^T$ . Она может замещать  $R^{-1}$  (или  $\tilde{R}^{-1}$ ) в памяти.

3. Перестановка строк и столбцов матрицы  $\tilde{R}^{-1}(\tilde{R}^{-1})^T$ . На каждом шаге матрица симметрична, поэтому в хранении нуждается лишь ее верхняя треугольная часть.

Чтобы вывести формулы обращения треугольной матрицы  $R$ , обозначим элементы  $R^{-1}$  через  $t_{ij}$ . Тогда из тождества  $R^{-1}R = I$  и того обстоятельства, что обе матрицы  $R$  и  $R^{-1}$  верхние треугольные, получаем уравнения

$$\sum_{l=i}^j t_{il} r_{lj} = \delta_{ij} = \begin{cases} 1, & i=j, \\ 0, & i \neq j, \end{cases}$$

$$1 \leq i \leq n, \quad i \leq j \leq n.$$

Решая относительно  $t_{ij}$ , находим

$$t_{ij} = \begin{cases} r_{jj}^{-1}, & j=i, \\ -r_{jj}^{-1} \sum_{l=i}^{j-1} t_{il} r_{lj}, & j>i. \end{cases}$$

Чтобы при вычислении элементов  $t_{ij}$  иметь возможность замещать ими элементы  $r_{ij}$  и заодно сократить число операций деления, эти формулы следует использовать таким образом:

$$t_{ii} = r_{ii}^{-1}, \quad i=1, \dots, n,$$

$$t_{ij} = -t_{ij} \sum_{l=i}^{j-1} t_{il} r_{lj}, \quad j=i+1, \dots, n, \quad i=1, \dots, n-1.$$

Счет по этим формулам требует  $n^3/6 + O(n^2)$  операций, где под операцией понимается "умножить и сложить" или "поделить и сложить". Умножение  $R^{-1}$  справа на транспонированную матрицу (см. (12.9)) также требует  $n^3/6 + O(n^2)$  операций. Следовательно, алгоритм COV вычисляет элементы верхнего треугольника матрицы  $(A^T A)^{-1}$  по элементам  $R$  за  $n^3/3 + O(n^2)$  операций.

В алгоритме COV предполагается, что задаваемая на входе матрица  $R$  (или  $\tilde{R}$ ) из правой части формулы (12.6) (или (12.7)) занимает верхнюю треугольную часть массива с именем  $A$ . На выходе в верхней треугольной части массива  $A$  будет находиться верхняя треугольная часть матрицы  $(A^T A)^{-1}$ . Массив с именем  $p$  содержит информацию (полученную, например, алгоритмом HFTI (см. (14.9)), описывающую действие матрицы  $A$  (из (12.10)).

Алгоритм 12.12. COV ( $A, n, p$ ):

1. Для  $i = 1, \dots, n$  положить  $a_{ii} := 1/a_{ii}$ .
2. Если  $n = 1$ , перейти к шагу 8.
3. Для  $i = 1, \dots, n-1$  выполнить нижеследующее до шага 7 включительно.
  4. Для  $j = i+1, \dots, n$  выполнить нижеследующее до шага 7 включительно.
    5. Положить  $s := 0$ .
    6. Для  $l = i, \dots, j-1$  положить  $s := s + a_{il} a_{lj}$ .
    7. Положить  $a_{ij} := -a_{ij} s$ .

8. Для  $i = 1, \dots, n$  выполнить нижеследующее до шага 12 включительно.

9. Для  $j = i, \dots, n$  выполнить нижеследующее до шага 12 включительно.

10. Положить  $s := 0$ .

11. Для  $l = j, \dots, n$  положить  $s := s + a_{il} a_{lj}$ .

12. Положить  $a_{ij} := s$ .

13. *Замечание.* Закончено вычисление элементов верхнего треугольника матрицы  $(A^T A)^{-1}$  для случая (12.3), где не делается перестановок при вычислении  $R$ . В случае же (12.4) нужно еще выполнить шаги 14–23, соответствующие левому умножению на  $P$  и правому умножению на  $P^T$  (см. (12.10)). В этом случае мы предполагаем наличие массива целых чисел  $p_i, i = 1, \dots, n$ . Смысл этих чисел такой:  $i$ -я перестановка в процессе приведения  $A$  к треугольному виду была перестановкой столбцов с номерами  $i$  и  $p_i$ .

14. Для  $i = n, \dots, 1$  выполнить нижеследующее до шага 22 включительно.

15. Если  $p_i = i$ , перейти к шагу 22.

16. Положить  $k := p_i$ . Поменять местами содержимое ячеек  $a_{ii}$  и  $a_{kk}$ . Если  $i = 1$ , перейти к шагу 18.

17. Для  $l = 1, \dots, i-1$  поменять местами содержимое ячеек  $a_{il}$  и  $a_{lk}$ .

18. Если  $k - i = 1$ , перейти к шагу 20.

19. Для  $l = i+1, \dots, k-1$  поменять местами содержимое ячеек  $a_{il}$  и  $a_{lk}$ .

20. Если  $k = n$ , перейти к шагу 22.

21. Для  $l = k+1, \dots, n$  поменять местами содержимое ячеек  $a_{il}$  и  $a_{kl}$ .

22. Continue.

23. *Замечание.* Вычисление нешкалированной ковариационной матрицы закончено.

Если используется сингулярное разложение (см. (12.5)), то для нешкалированной ковариационной матрицы  $C$  справедлива формула (12.11). Поэтому индивидуальные элементы  $C$  выражаются формулами

$$c_{ij} = \sum_{k=1}^n \frac{v_{ik} v_{jk}}{s_{kk}^2}. \quad (12.13)$$

Если матрица  $V$  вычислена в явном виде и хранится массивом размера  $n \times n$ , то на ее место можно записать матрицу  $V S^{-1}$ . Вслед за этим верхний треугольник  $V S^{-1}$  может быть замещен верхней треугольной частью  $C$ . Это требует использования дополнительного массива длины  $n$ .

*Замечания о некоторых альтернативах вычисления  $C$*

Мы хотим указать три стандартных случая, когда вычисляется нешкалированная ковариационная матрица  $C$ .

1. Матрица  $C$  (или  $C^{-1}$ ) может встречаться в качестве промежуточной величины в некоторой статистической формуле.

2. Матрица  $C$  может быть выведена на печать (или, например, экран дисплея), чтобы пользователь (заказчик задачи) мог исследовать и интерпретировать ее.



3. Какое-то подмножество элементов матрицы  $C$  может использоваться внутри некоторого автоматизированного процесса в качестве управляющих параметров.

В каждом из этих случаев мы укажем иной подход, который может оказаться более информативным или более экономичным.

Рассмотрим вначале случай 1. Обычно формулы, включающие в себя  $C$  или  $C^{-1}$ , содержат выражения вида

$$E = BCB^T \quad (12.14)$$

или

$$F = D^T C^{-1} D. \quad (12.15)$$

Наиболее эффективным и численно устойчивым методом вычисления таких выражений обычно является использование факторизованной формы  $C$  или  $C^{-1}$ :

$$C = (R^{-1})(R^{-1})^T \quad (12.16)$$

или

$$C^{-1} = R^T R \quad (12.17)$$

соответственно.

В этом случае произведение (12.14) вычисляется так: вначале решается относительно  $Y$  уравнение  $YR = B$ , а затем строится  $E = YY^T$ .

Аналогично в случае (12.15) вычисляется  $X = RD$ , а затем  $F = X^T X$ .

Важный общий принцип, иллюстрируемый этими замечаниями, таков: если имеется факторизация (12.16) матрицы  $C$  или факторизация (12.17) матрицы  $C^{-1}$ , то можно на основе этих факторизаций построить вычислительные процедуры, более экономичные и более устойчивые, чем те, что требуют явного вычисления  $C$  или  $C^{-1}$ .

Эти замечания легко модифицировать на тот случай, когда в выражениях для матриц  $C$  или  $C^{-1}$  входит матрица перестановки  $P$  (см. формулы (12.7) и (12.10) соответственно).

В случае 2 пользователь часто исследует корреляции (или почти зависимости) между различными парами компонент решения  $x$ . В этом контексте принято строить вспомогательную матрицу

$$E = D^{-1} C D^{-1}, \quad (12.18)$$

где  $D$  — диагональная матрица,  $i$ -й диагональный элемент  $d_{ii}$  которой равен  $c_{ii}^{1/2}$ . Тогда  $e_{ii} = 1$ ,  $i = 1, \dots, n$ , и  $|e_{ij}| \leq 1$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, n$ .

Наличие элемента  $e_{ij}$ ,  $i \neq j$ , близкого к 1 или -1 (например,  $e_{ij} = 0,95$ ), воспринимается как указание на то, что  $i$ -я и  $j$ -я компоненты решения сильно коррелированы. Алгебраически это соответствует тому, что главная  $2 \times 2$ -подматрица

$$\begin{bmatrix} e_{ii} & e_{ij} \\ e_{ji} & e_{jj} \end{bmatrix}$$

почти вырождена.

Слабость анализа этого типа состоит в том, что таким путем легко обнаружить только зависимости между парами переменных. В то же время может иметься, например, почти зависимая группа из трех переменных, в которой никакие две переменные не являются почти зависимыми. Такой

группе из трех переменных можно сопоставить главную  $3 \times 3$ -подматрицу матрицы  $E$ , например

$$\begin{bmatrix} 1 & -0,49 & -0,49 \\ -0,49 & 1 & -0,49 \\ -0,49 & -0,49 & 1 \end{bmatrix}. \quad (12.19)$$

Здесь никакой внедиагональный элемент не близок к 1 или -1 и, следовательно, нет почти вырожденной главной  $2 \times 2$ -подматрицы. В то же время  $3 \times 3$ -подматрица почти вырождена. Она становится вырожденной точно, если значения внедиагональных элементов изменить с -0,49 на -0,50.

Зависимости между тремя или большим числом переменных очень трудно обнаружить визуальным исследованием матриц  $C$  или  $E$ . Такие зависимости, однако, выявляет матрица  $V$  сингулярного разложения  $A = USV^T$  (или, что эквивалентно, спектрального разложения  $A^T A = VS^2 V^T$ ).

Пусть  $s_j$  — сингулярное число  $A$ , малое в сравнении с наибольшим сингулярным числом  $s_1$ . Соответствующие столбцы  $u_j$  и  $v_j$  матриц  $V$  и  $U$  связаны равенством

$$A v_j = s_j u_j, \quad (12.20)$$

откуда

$$\|A v_j\| = s_j. \quad (12.21)$$

Малость отношения  $s_j/s_1$  может быть принята в качестве критерия почти вырожденности матрицы  $A$ ; при этом вектор  $u_j$  указывает конкретную линейную комбинацию столбцов  $A$ , почти равную нулю. Визуальное исследование столбцов  $V$ , отвечающих малым сингулярным числам, оказалось очень полезным приемом анализа плохо обусловленных задач наименьших квадратов.

В случае 3 следует заметить, что нет необходимости вычислять всю матрицу  $C$ , если требуется лишь некоторое подмножество ее элементов. Если, например, уже построена матрица  $R^{-1}$ , то отдельные элементы  $C$  можно вычислять независимо по формулам, выводимым непосредственно из (12.9). В типичном случае нужны лишь некоторые диагональные элементы  $C$  или какая-либо ее главная подматрица.

## ГЛАВА 13

### ВЫЧИСЛЕНИЕ РЕШЕНИЯ НЕООПРЕДЕЛЕННОЙ ЗАДАЧИ ПОЛНОГО РАНГА

Рассмотрим теперь задачу НК  $Ax = b$ , где  $m < n$ ,  $\text{rang } A = m$  (случай 3а на рис. 1.1). Алгоритм ее решения может быть разбит на следующие шаги:

$$AQ = [R: 0]; \quad (13.1)$$

$$R y_1 = b; \quad (13.2)$$

$$y_2 \text{ — произвольный вектор}; \quad (13.3)$$

$$x = Q y = Q \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \quad (13.4)$$

Алгоритм вычислит ортогональную матрицу  $Q$  из (13.1), при этом  $R$  будет невырожденной нижней треугольной матрицей. Существование обеих матриц было установлено путем транспонирования основного разложения теоремы 3.11.

В (13.2)  $m$ -вектор  $y_1$  определен однозначно. При любом  $n - m$ -векторе  $y_2$  (см. (13.3)) вектор  $x$  из (13.4) удовлетворяет уравнению  $Ax = b$ . Решение минимальной длины (нормальное решение) получается в соответствии с теоремой 2.3 при  $y_2 = 0$ .

Существуют ситуации, когда недоопределенная задача  $Ax = b$  является частью большей оптимизационной задачи: при этом  $y_2$  оказывается ненулевым вектором. Такова, в частности, задача минимизации  $\|Ex - f\|$  при наличии дополнительных линейных ограничений  $Ax = b$  (которые не определяют  $x$  однозначно). Эта задача, которую мы называем НКУ\*, будет рассмотрена в гл. 20–22.

Матрица  $Q$  из (13.1) строится как произведение

$$Q = Q_1 \dots Q_m, \quad (13.5)$$

где каждая  $Q_j$  имеет вид

$$Q_j = I_n + b_j^{-1} u^{(j)} u^{(j)T}, \quad j = 1, \dots, m. \quad (13.6)$$

Величина  $b_j$ , участвующая в (13.6), перевычисляется всякий раз, как она нужна, по формуле (10.10). Вводя индексы, можем написать

$$b_j = s_j u_j^{(j)}, \quad j = 1, \dots, m. \quad (13.7)$$

Каждая величина  $s_j$  — это  $j$ -й диагональный элемент матрицы  $R$  и будет храниться как таковой. Описываемый алгоритм будем называть НВТ ( $m, n, A, g$ ). Входная информация алгоритма НВТ состоит из целых чисел  $m$  и  $n$  и  $m \times n$ -матрицы  $A$  ранга  $m$ . На выходе будет получена нижняя треугольная матрица  $R$ , хранимая в нижней треугольной части массива с именем  $A$ . Величины  $u_j^{(j)}$  будут находиться во вспомогательном массиве переменных с именами  $g_j, j = 1, \dots, m$ . Остальные ненулевые компоненты вектора  $u^{(j)}$  хранятся в наддиагональной части  $j$ -й строки массива  $A$ .

Алгоритм 13.8. НВТ ( $m, n, A, g$ ):

1. Для  $j := 1, \dots, m$  выполнить алгоритм Н1 ( $j, j + 1, n, a_{j1}, g_j, a_{j+1,1}, m - j$ ) (см. (10.22)).

2. Комментарий. Приведение к (нижнему) треугольному виду закончено.

По поводу шага 1 следует отметить, что  $a_{j1}$  указывает начало в памяти вектор-строки, а  $a_{j+1,1}$  — начало подматрицы, образованной  $m - j$  строками, к которым применяется преобразование.

Приводимый ниже алгоритм выполняет вычисления в соответствии с формулами (13.2)–(13.4). Входная информация этого алгоритма состоит из целых чисел  $m$  и  $n$  и массивов с именами  $A$  и  $g$  в том виде, как они получены на выходе алгоритма НВТ (см. 13.8). Решение будет помещено в массив с именем  $x$ . Вектор  $b$  будет замещен в памяти вектором  $y_1$  (см. (13.2)).

Алгоритм 13.9. HS2 ( $m, n, A, g, b, x$ ):

1. Положить  $b_1 := b_1 / a_{11}$ .

2. Для  $i := 2, \dots, m$  положить  $b_i = \frac{1}{a_{ii}} (b_i - \sum_{j=1}^{i-1} a_{ij} b_j)$ .

3. Комментарий. Здесь следует задать вектор  $y_2$ . Если требуется нормальное решение, нужно положить  $y_2 := 0$ .

4. Положить  $x := \begin{bmatrix} b \\ y_2 \end{bmatrix}$ .

5. Для  $j := m, m - 1, \dots, 1$  выполнить алгоритм Н2 ( $j, j + 1, n, a_{j1}, g_j, x, 1$ ).

6. Комментарий. Массив с именем  $x$  содержит теперь частное решение системы  $Ax = b$ . Если на шаге 4 в качестве  $y_2$  был взят нулевой вектор, то полученное решение является нормальным (т.е. имеет минимальную длину).

## ГЛАВА 14

### ВЫЧИСЛЕНИЕ РЕШЕНИЯ ЗАДАЧИ НК, ВОЗМОЖНО, НЕПОЛНОГО ПСЕВДОРАНГА

Алгоритмы гл. 11–13 для задач полного ранга ориентированы на те случаи, когда матрица  $A$  предполагается достаточно хорошо обусловленной, т.е. исключена возможность, что неопределенность входных данных или возмущение  $A$ , эквивалентное округлениям в процессе вычислений, приведут к замене  $A$  матрицей неполного ранга. Оценки для возмущений второго типа будут получены в гл. 15–17.

Есть, однако, и другие случаи, когда желательно иметь алгоритм, определяющий, насколько близка данная матрица к матрице неполного ранга. Если выяснилось, что при изменениях коэффициентов исходной матрицы, имеющих порядок неопределенности ее задания, может получиться матрица неполного ранга, то алгоритм должен принять соответствующие меры.

Алгоритм должен по меньшей мере известить пользователя при обнаружении такой ситуации неполного (в указанном смысле) ранга. Вдобавок пользователь может пожелать, чтобы решение было вычислено посредством процедуры, свободной от произвольных возмущений, сопутствующих вычислениям, в которых очень плохо обусловленная задача обрабатывается как задача полного ранга.

Один из методов стабилизации такой задачи состоит в замене  $A$  близкой матрицей неполного ранга (назовем ее  $\tilde{A}$ ) и последующем вычислении нормального псевдорешения задачи  $\tilde{A}x \cong b$ . Эта замена обычно производится неявно как часть численного алгоритма. Алгоритм НФТИ, описываемый в данной главе, принадлежит к алгоритмам этого типа. Пример, иллюстрирующий использование алгоритма НФТИ и некоторых других методов стабилизации, приведен в гл. 26. В гл. 25 описаны другие методы стабилизации.

Определим псевдоранг  $k$  матрицы  $A$  как ранг матрицы  $\tilde{A}$ , заменяющей  $A$  в результате выполнения конкретного вычислительного алгоритма. Заметим, что псевдоранг не является функцией одной лишь матрицы  $A$ , но зависит также от других факторов: деталей численного алгоритма; значений допусков, используемых в вычислениях; влияния округлений и т.д.

\* В оригинале — LSE (Least Squares with Equality constraints). (Примеч. пер.)

Алгоритм HFTI применим, в частности, к задачам неполного ранга указанным на рис. 1.1 как случаи 1b, 2b и 3b. Строго говоря, однако, имеет не псевдоранг, а ранг  $A$  будет определен в ходе вычислений.

Исходя из теорем 2.3 и 3.19, сформулируем математические соотношения, лежащие в основе алгоритма HFTI:

$$QAP = R \equiv \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix} \quad (14.1)$$

$$Qb = c \equiv \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \begin{matrix} k \\ m-k \end{matrix} \quad (14.2)$$

$$[R_{11} : R_{12}] K = [W : 0], \quad (14.3)$$

$$Wy_1 = c_1, \quad (14.4)$$

$$y_2 - \text{произвольный вектор}, \quad (14.5)$$

$$x = PK \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \equiv PKy, \quad (14.6)$$

$$\|b - Ax\| = \|c_2 - R_{22}y_2\| (= \|c_2\| \text{ при } y_2 = 0). \quad (14.7)$$

Алгоритм определит ортогональную матрицу  $Q$  и матрицу перестановки  $P$  такие, что  $R$  будет верхней треугольной матрицей с невырожденной подматрицей  $R_{11}$ .

Матрица перестановки  $P$  строится неявно как результат стратегии перестановок столбцов, принятой в алгоритме. Эта стратегия тесно связана с задачей определения псевдоранга  $k$ . Существенна невырожденность подматрицы  $R_{11}$ . Если специальные требования конкретных приложений не заставляют выбрать иной критерий, то предпочтительным решением будет достаточно хорошо обусловленная матрица  $R_{11}$  и матрица  $R_{22}$  с достаточной малой нормой.

Пример такой нестандартной ситуации — случай, когда заранее известно, что  $A_{n \times n}$  имеет простое собственное значение нуль. В этом случае можно было бы положить  $k = n - 1$  вместо того, чтобы определять значение  $k$  посредством алгоритма.

Другая нестандартная ситуация — это задача НК с весами (гл. 22), где может допускаться очень плохо обусловленная матрица  $R_{11}$ .

В алгоритме HFTI используется следующая стратегия перестановок столбцов. Перед построением  $j$ -го преобразования Хаусхолдера среди столбцов с номерами  $j, \dots, n$  выбирается тот (назовем его  $\lambda$ ), у которого сумма квадратов компонент в строках  $s$   $j$ -й по  $m$ -ю наибольшая. Содержимое столбцов  $j$  и  $\lambda$  затем меняется местами, после чего строится  $j$ -е преобразование Хаусхолдера так, чтобы аннулировать элементы в позициях  $aj$ ,  $i = j + 1, \dots, m$ .

Некоторой экономии счетного времени можно добиться, если перестраивать от шага к шагу нужные суммы квадратов\*). Это возможно благо-

\*)  $A$  не вычислять их заново на каждом шаге. (Примеч. пер.)

даря ортогональности преобразования Хаусхолдера. Подробности перестройки даны в описании алгоритма HFTI (шаги 3–10).

В результате принятой стратегии перестановок диагональные элементы  $R$  не будут возрастать по абсолютной величине. Более того, они будут удовлетворять неравенствам (6.16). См. в этой связи теоремы 6.13 и 6.31.

В алгоритме HFTI псевдоранг  $k$  определяется как наибольший индекс  $j$ , для которого  $|r_{jj}| > \tau$ , где  $\tau$  — задаваемый пользователем неотрицательный параметр. Стратегия перестановок столбцов и подходящий выбор  $\tau$ , разумеется, зависят от исходного масштабирования матрицы. Этот предмет обсуждается в гл. 25.

Выбор значения  $k$  меньшего, чем  $\min(m, n)$ , равносителен замене заданной матрицы

$$A = Q^T \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} P^T$$

матрицей

$$\tilde{A} = Q^T \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} P^T \quad (14.8)$$

Заметим, что  $\|\tilde{A} - A\| = \|R_{22}\|$ ,  $\|\tilde{A} - A\|_F = \|R_{22}\|_F$ . Далее при указанных процедурах перестановки столбцов и определения псевдоранга имеем

$$\|R_{22}\| \leq (\mu - k)^{1/2} |r_{k+1, k+1}| \leq (\mu - k)^{1/2} \tau,$$

$$\mu = \min(m, n).$$

Ортогональная матрица  $K$  в (14.3) выбирается так, чтобы  $W$  была невырожденной верхней треугольной  $k \times k$ -матрицей;  $k$ -вектор  $y_1$  вычисляется как единственное решение (14.4). Для  $n - k$ -вектора  $y_2$  можно задать произвольное значение; при этом нулевому  $y_2$  соответствует решение с минимальной длиной. Окончательно решение  $x$  выражается формулой (14.6). Норму невязки можно вычислить как правую часть формулы (14.7).

На входе в алгоритм HFTI задаются матрица  $A$  и вектор  $b$ , хранимые в одноименных массивах, целые числа  $m, n$  и неотрицательный параметр  $\tau$ . Ортогональная матрица  $Q$  из формулы (14.1) является произведением  $\mu$  преобразований Хаусхолдера  $Q_i$ . Информация, определяющая эти матрицы, занимает на выходе нижнюю треугольную часть массива  $A$  плюс  $\mu$  дополнительных ячеек массива с именем  $h$ .

Массив  $h$  используется также для хранения квадратов длин столбцов некоторых подматриц, генерируемых в процессе вычислений. Эти числа нужны при определении перестановок столбцов. На их место можно записать (и действительно записываются) главные элементы матриц  $Q_i$ .

Матрица перестановки  $P$  строится как произведение матриц-транспозиций  $P = (1, p_1) \dots (\mu, p_\mu)$ . Здесь  $(i, j)$  обозначает матрицу перестановки, получаемую путем транспозиции столбцов  $i$  и  $j$  единичной матрицы  $I_n$ . Целые числа  $p_i$  записываются в массив  $p$ . Ортогональная матрица  $K$  из (14.3) есть произведение  $k$  преобразований Хаусхолдера  $K_i$ . Информация, определяющая эти матрицы, занимает прямоугольную часть массива  $A$ , образованную первыми  $k$  строками и последними  $n - k$  столбцами, и еще  $k$  дополнительных ячеек в массиве  $g$ .

$A:$

|          |          |          |           |           |
|----------|----------|----------|-----------|-----------|
| $w_{11}$ | $w_{12}$ | $w_{13}$ | $k_{14}$  | $k_{15}$  |
| $u_{12}$ | $w_{22}$ | $w_{23}$ | $k_{24}$  | $k_{25}$  |
| $u_{13}$ | $u_{23}$ | $w_{33}$ | $k_{34}$  | $k_{35}$  |
| $u_{14}$ | $u_{24}$ | $u_{34}$ | $\dagger$ | $\dagger$ |
| $u_{15}$ | $u_{25}$ | $u_{35}$ | $u_{45}$  | $\dagger$ |
| $u_{16}$ | $u_{26}$ | $u_{36}$ | $u_{46}$  | $u_{56}$  |

Массив, в котором первоначально хранится матрица  $A$ . По окончании работы в нем хранятся матрица  $W$  и большая часть информации, необходимой для построения ортогональных преобразований Хаусхолдера

$h:$

|          |          |          |          |          |
|----------|----------|----------|----------|----------|
| $u_{11}$ | $u_{22}$ | $u_{33}$ | $u_{44}$ | $u_{55}$ |
|----------|----------|----------|----------|----------|

Главные элементы преобразований Хаусхолдера  $Q_i$

$g:$

|          |          |          |           |           |
|----------|----------|----------|-----------|-----------|
| $k_{11}$ | $k_{22}$ | $k_{33}$ | $\diagup$ | $\diagup$ |
|----------|----------|----------|-----------|-----------|

Главные элементы преобразований Хаусхолдера  $K_i$

$p:$

|       |       |       |       |       |
|-------|-------|-------|-------|-------|
| $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|-------|-------|-------|-------|-------|

Информация о перестановках

Рис. 14.1

На рис. 14.1 показано распределение памяти на выходе для  $m = 6$ ,  $n = 5$  и  $k = \text{rank } A = 3$ . Если требуется сохранить полную информацию об исходной матрице  $A$ , то нужно использовать дополнительный массив для записи диагональных элементов матрицы  $R_{11}$  из формулы (14.1). Они понадобятся для вычисления матрицы  $Q$  той же формулы. Так как в алгоритме HFTI матрица  $Q$  применяется непосредственно к вектору  $b$ , то здесь этого дополнительного массива не нужно.

По окончании работы алгоритма вектор решения  $x$  находится в одноименном массиве, а вектор  $c$  — в массиве с именем  $b$ . Алгоритм организован таким образом, что имена  $b$  и  $x$  могут отвечать одному и тому же массиву памяти, т.е. нет необходимости в дополнительном массиве для хранения  $x$ . В этом случае длина массива  $b$  должна быть равна  $\max(m, n)$ .

Шаги 1–13 алгоритма HFTI описывают приведение к верхнему треугольному виду с использованием перестановок столбцов. В сущности, это — алгоритм, предложенный в [72]. Дополнительные шаги соответствуют обобщению алгоритма на случай неполного ранга [90].

В формулировке шагов 3–10 мы следовали деталям программы, содержащейся в [25]. Проверка на шаге 6, по существу, эквивалентна условию: "Если  $h_\lambda > 10^3 \eta h$ ". Форма этого условия, которую мы используем, позволяет избежать явного указания машинно зависимого параметра  $\eta$ . Его смысл — относительная точность машинной арифметики.

Алгоритм 14.9. HFTI ( $A, m, n, b, \tau, x, k, h, g, p$ ):

1. Положить  $\mu := \min(m, n)$ .
2. Для  $j = 1, \dots, \mu$  выполнить шаги 3–12.
3. Если  $j = 1$ , перейти к шагу 7.
4. Для  $l = j, \dots, n$  положить  $h_l := h_l - a_{j-1, l}^2$ .

5. Определить  $\lambda$  такое, что  $h_\lambda := \max \{ h_l : j < l \leq n \}$ .

6. Если  $(h + 10^{-3} h_\lambda) > h$ , перейти к шагу 9.

7. Для  $l = j, \dots, n$  положить  $h_l := \sum_{i=j}^m a_{li}^2$ .

8. Определить  $\lambda$  такое, что  $h_\lambda := \max \{ h_l : j < l \leq n \}$ . Положить  $\bar{h} := h_\lambda$ .

9. Положить  $p_j := \lambda$ . Если  $p_j = j$ , перейти к шагу 11.

10. Переставить столбцы  $j$  и  $\lambda$  матрицы  $A$  и положить  $h_\lambda := h_j$ .

11. Выполнить алгоритм H1( $j, j+1, m, a_{1j}, h_j, a_{1, j+1}, n-j$ ).

12. Выполнить алгоритм H2( $j, j+1, m, a_{1j}, h_j, b, 1$ ).

13. Комментарий. Теперь следует определить псевдоранг  $k$ . Заметим, что диагональные элементы  $R$  (храняемые в позициях  $a_{11}, \dots, a_{\mu\mu}$ ) не возрастают по абсолютной величине. HFTI выбирает в качестве  $k$  наибольший индекс  $j$  такой, что  $|a_{jj}| > \tau$ . Если  $|a_{jj}| \leq \tau$  для всех  $j$ , то псевдорангу  $k$  присваивается значение 0, вектор решения  $x$  полагается равным нулю и алгоритм заканчивается.

14. Если  $k = n$ , перейти к шагу 17.

15. Комментарий. Здесь  $k < n$ . Сейчас будут определены ортогональные преобразования  $K_i$ , чье произведение составляет матрицу  $K$  формулы (14.3).

16. Для  $i = k, k-1, \dots, 1$  выполнить алгоритм H1( $i, k+1, n, a_{i1}, g_i, a_{11}, i-1$ ). (Параметры  $a_{i1}$  и  $a_{11}$  указывают первые элементы соответствующих строчных векторов.)

17. Положить  $x_k := b_k / a_{kk}$ . Если  $k \leq 1$ , перейти к шагу 19.

18. Для  $i = k-1, k-2, \dots, 1$  положить  $x_i := \frac{1}{a_{ii}} \left( b_i - \sum_{j=i+1}^k a_{ij} x_j \right)$  (см. (14.4)).

19. Если  $k = n$ , перейти к шагу 22.

20. Задать  $n-k$ -вектор  $y_2$  (см. (14.5)) и поместить его компоненты в позиции  $x_i, i = k+1, \dots, n$ . В частности, если требуется решение минимальной длины, нужно положить  $y_2 := 0$ .

21. Для  $i = 1, \dots, k$  выполнить алгоритм H2( $i, k+1, n, a_{i1}, g_i, x, 1$ ) (см. (14.6). Параметр  $a_{i1}$  указывает первый элемент строчного вектора).

22. Для  $j = \mu, \mu-1, \dots, 1$  выполнить шаг 23.

23. Если  $p_j \neq j$ , переставить содержимое ячеек  $x_j$  и  $x_{p_j}$  (см. (14.6)).

## ГЛАВА 15

### АНАЛИЗ ПОГРЕШНОСТЕЙ ОКРУГЛЕНИЙ ДЛЯ ПРЕОБРАЗОВАНИЙ ХАУСХОЛДЕРА

В предыдущих главах мы рассматривали преимущественно математические аспекты различных задач метода наименьших квадратов. В гл. 9 было исследовано воздействие на решение неопределенности во входных данных задачи. Практические машинные вычисления связаны с погрешностями округлений. Замечательный вывод теории погрешностей округлений [7] состоит в том, что для многих классов задач линейной

алгебры вычисленное машинное решение является точным для задачи, получаемой из исходной сравнительно малым возмущением коэффициентов.

Большое значение имеет тот факт, что возмущения коэффициентов, эквивалентные произведенным округлениям, часто бывают меньше, чем неопределенность задания этих коэффициентов; в таких случаях погрешности округлений с практической точки зрения можно игнорировать.

Цель гл. 15–17 представить результаты этого типа для алгоритмов наименьших квадратов, описанных в гл. 11–14.

Действительные числа, точно представимые на данной машине в виде нормализованных чисел с плавающей запятой, будем называть машинными числами.

Емкость арифметики с плавающей запятой данной машины удобно характеризовать тремя положительными числами  $L, U, \eta$ . Число  $L$  — это наименьшее положительное число такое, что и  $L$ , и  $-L$  являются машинными числами. Число  $U$  — это наибольшее положительное число с тем свойством, что и  $U$ , и  $-U$  являются машинными числами.

В последующем анализе предполагается, что все ненулевые числа  $x$ , участвующие в вычислениях, удовлетворяют условиям  $L \leq |x| \leq U$ . Надежная универсальная программа, реализующая описываемые алгоритмы, может обеспечить неравенство  $|x| \leq U$  путем соответствующего масштабирования. При этом может оказаться невозможным одновременно гарантировать условие  $|x| \geq L$  для всех ненулевых чисел  $x$ . В то же время подходящим масштабированием можно добиться того, чтобы числа, для которых  $|x| < L$ , можно было заменить нулями, не влияя сколько-нибудь существенно на точность результатов в смысле векторных норм.

Число  $\eta$  характеризует относительную точность машинной арифметики с плавающей запятой и нормализацией. Следуя Уилкинсону, мы используем обозначение

$$\bar{z} = \text{fl}(x + y)$$

для указания на то, что  $\bar{z}$  есть машинное число, полученное в результате машинной операции сложения машинных чисел  $x$  и  $y$ . Если  $z = x + y$ , то даже при выполнении условий  $L \leq |z| \leq U$  часто оказывается, что  $\bar{z} - z \neq 0$ . Удобно называть эту разность погрешностью округления независимо от того, действительно ли округляет данная машина или же оно просто производит усечение \*).

Число  $\eta$  — это наименьшее положительное число с таким свойством: если точный результат  $z$  одной из пяти операций — сложить, вычесть, умножить, разделить, извлечь квадратный корень — равен нулю или удовлетворяет неравенствам  $L \leq |z| \leq U$ , то найдутся числа  $\epsilon_i$ , ограниченные по абсолютной величине числом  $\eta$ , для которых выполняется одно из следующих пяти уравнений:

$$\text{fl}(x + y) = x(1 + \epsilon_1) + y(1 + \epsilon_2) = \frac{x}{1 + \epsilon_3} + \frac{y}{1 + \epsilon_4} \quad (15.1)$$

(если  $x$  и  $y$  имеют одинаковые знаки, то  $\epsilon_1 = \epsilon_2$  и  $\epsilon_2 = \epsilon_4$ );

$$\text{fl}(x - y) = x(1 + \epsilon_5) - y(1 + \epsilon_6) = \frac{x}{1 + \epsilon_7} - \frac{y}{1 + \epsilon_8} \quad (15.2)$$

\* В оригинале — truncation. (Примеч. пер.)

(если  $x$  и  $y$  имеют противоположные знаки, то  $\epsilon_5 = \epsilon_6$  и  $\epsilon_7 = \epsilon_8$ );

$$\text{fl}(xy) = xy(1 + \epsilon_9) = \frac{xy}{1 + \epsilon_{10}} \quad (15.3)$$

$$\text{fl}\left(\frac{x}{y}\right) = \left(\frac{x}{y}\right)(1 + \epsilon_{11}) = \frac{x}{y(1 + \epsilon_{12})} \quad (15.4)$$

$$\text{fl}(x^{1/2}) = x^{1/2}(1 + \epsilon_{13}) = \frac{x^{1/2}}{1 + \epsilon_{14}} \quad (15.5)$$

Общий подход к анализу алгоритмов линейной алгебры на основе предположений этого типа хорошо отражен в литературе [7, 11, 23, 24, 190]. При этом использовались различные приемы, позволяющие учесть неизбежно возникающие члены второго и более высоких порядков по  $\eta$ . Так как в оценках наиболее полезную информацию дает коэффициент при  $\eta$ , то только он и будет определяться в последующем анализе.

Вследствие группировки членов порядка  $\eta^2$  в единое слагаемое  $O(\eta^2)$  "оценки", которые будут получены в гл. 15–17, не являются вполне вычислимыми. Однако для значений  $\eta$ , соответствующих реальным машинам, этим слагаемым можно пренебречь. Так как повсюду в анализе предполагается максимальное накопление погрешностей, то выводимые нами оценки существенно превышают реальные погрешности. Основное практическое назначение этих оценок — показать зависимость погрешностей от параметров  $m, n, \eta$  и установить численную устойчивость алгоритмов из гл. 11, 13, 14.

Вначале мы проанализируем погрешности округлений в алгоритме H1 (см. 10.22). Будем считать, что алгоритм строит  $m \times m$ -преобразование Хаусхолдера, которое аннулирует все элементы данного  $m$ -вектора  $v$ , кроме первого. Математически алгоритм определяется следующими формулами:

$$\mu = \max_i |v_i| \quad (15.6)$$

$$t = \sum_{i=1}^m \left( \frac{v_i}{\mu} \right)^2 \quad (15.7)$$

$$s = -(\text{sgn } v_1) t^{1/2} \mu \quad (15.8)$$

$$u_1 = v_1 - s \quad (15.9)$$

$$u_i = v_i, \quad i = 2, \dots, m, \quad (15.10)$$

$$b = su_1, \quad (15.11)$$

$$Q = I_m + b^{-1}uu^T. \quad (15.12)$$

В алгоритм включено масштабирование, имеющее целью избежать машинных нулей. Условимся черточкой сверху отмечать величины, являющиеся машинными числами или составленные из таких чисел. Величины, реально

вычисленные алгоритмом, можно представить формулами

$$\mu = \max_i |v_i|, \quad (15.13)$$

$$\bar{r} = \text{fl} \left[ \sum_{i=1}^m \left( \frac{v_i}{\mu} \right)^2 \right], \quad (15.14)$$

$$\bar{s} = -\text{fl}[(\text{sgn}(v_1))(\bar{r})^{1/2} \mu], \quad (15.15)$$

$$\bar{u} = \text{fl}(v_1 - \bar{s}), \quad (15.16)$$

$$\bar{u}_i = v_i, \quad i = 2, \dots, m, \quad (15.17)$$

$$\bar{b} = \text{fl}(\bar{s} \bar{u}_1), \quad (15.18)$$

$$\bar{Q} = \text{fl}(I_m + \bar{b}^{-1} \bar{u} \bar{u}^T). \quad (15.19)$$

Вычисленные величины, выражаемые формулами (15.14)–(15.19), связаны с одноименными точными величинами из формул (15.7)–(15.12) следующим образом:

$$\bar{r} = r(1 + \tau), \quad |\tau| \leq (m+2)\eta + O(\eta^2), \quad (15.20)$$

$$\bar{s} = s(1 + \sigma), \quad |\sigma| \leq \frac{(m+6)\eta}{2} + O(\eta^2), \quad (15.21)$$

$$\bar{u}_1 = u_1(1 + \kappa), \quad |\kappa| \leq \frac{(m+8)\eta}{2} + O(\eta^2), \quad (15.22)$$

$$\bar{u}_i = u_i, \quad i = 2, \dots, m, \quad (15.23)$$

$$\bar{b} = b(1 + \beta), \quad |\beta| \leq (m+8)\eta + O(\eta^2), \quad (15.24)$$

$$\|\bar{Q} - Q\|_F \leq 4\kappa + 2\beta + O(\eta^2) \leq (4m+32)\eta + O(\eta^2). \quad (15.25)$$

Мы опускаем громоздкий вывод этих оценок, поскольку он очень схож с выводом оценок книги [7]. Заметим, однако, что наши оценки отличаются от оценок Уилкинсона. Главная причина различия следующая: мы предполагаем, что вся арифметика выполняется с точностью  $\eta$ , а Уилкинсон считает, что некоторые операции выполняются с точностью  $\eta^2$ . Мы отложим обсуждение такой арифметики со смешанной точностью до гл. 17.

Сейчас мы рассмотрим применение преобразования Хаусхолдера к  $m \times n$ -матрице  $C$  посредством алгоритма H2 или второй (необязательной) части алгоритма H1 (см. 10.22). С математической точки зрения нужно вычислить произведение

$$B = QC. \quad (15.26)$$

При реальных вычислениях, однако, мы имеем лишь  $\bar{Q}$  вместо  $Q$ ; поэтому будет вычислена матрица

$$\bar{B} = \text{fl}(\bar{Q}C), \quad (15.27)$$

$$\|\bar{B} - B\|_F = \|\text{fl}(\bar{Q}C) - \bar{Q}C\|_F + \|\bar{Q}C - QC\|_F \leq$$

$$\leq \|\text{fl}(\bar{Q}C) - \bar{Q}C\|_F + \|\bar{Q} - Q\|_F \|C\|_F \equiv$$

$$\equiv \|E\|_F + \|\bar{Q} - Q\|_F \|C\|_F. \quad (15.28)$$

Подобно тому как это сделано в книге [7], можно показать, что

$$\|E\|_F \leq (2m+5) \|C\|_F \eta + O(\eta^2). \quad (15.29)$$

Теперь из (15.25), (15.28) и (15.29) получаем

$$\|\bar{B} - B\|_F \leq (6m+37) \|C\|_F \eta + O(\eta^2). \quad (15.30)$$

Мы должны теперь исследовать погрешность, связанную с применением к  $m \times n$ -матрице  $A$  к последовательных преобразований Хаусхолдера. Точный математический процесс описывается формулами

$$A_1 = A, \quad (15.31)$$

$$A_{i+1} = \hat{Q}_i A_i, \quad i = 1, \dots, k, \quad (15.32)$$

где матрица Хаусхолдера  $\hat{Q}_i$  аннулирует элементы  $i+1, \dots, m$  столбца  $i$  матрицы  $\hat{Q}_i A_i$ . Эта последовательность умножений лежит в основе различных алгоритмов, которые мы рассматриваем.

Будут вычислены величины

$$\bar{A}_1 = A, \quad (15.33)$$

$$\bar{A}_{i+1} = \text{fl}(\bar{Q}_i \bar{A}_i), \quad i = 1, \dots, k, \quad (15.34)$$

где  $\bar{Q}_i$  – численное приближение к точной матрице  $Q_i$ , которая аннулирует элементы  $i+1, \dots, m$  столбца  $i$  произведения  $Q_i A_i$ .

Определим матрицу погрешности

$$F_i = Q_i \bar{A}_i - \text{fl}(\bar{Q}_i \bar{A}_i) = Q_i \bar{A}_i - \bar{A}_{i+1}, \quad i = 1, \dots, k. \quad (15.35)$$

Из (15.30) следует оценка

$$\|F_i\|_F \leq \alpha_{m+1-i} \|\bar{A}_i\|_F + O(\eta^2), \quad (15.36)$$

где

$$\alpha_j = (6j+37)\eta. \quad (15.37)$$

Из ортогональности матриц  $Q_i$  прямой подстановкой выводится

$$\|\bar{A}_{k+1} - Q_k \dots Q_1 A\|_F \leq \sum_{i=1}^k \alpha_{m+1-i} \|\bar{A}_i\|_F + O(\eta^2) \leq$$

$$\leq \|A\|_F \sum_{i=1}^k \alpha_{m+1-i} + O(\eta^2) \leq$$

$$\leq (6m-3k+40)k\eta + O(\eta^2). \quad (15.38)$$

Полезно ввести также другую матрицу погрешности  $H_k = Q_1 \dots Q_k \bar{A}_{k+1} - A$ . Тогда

$$\bar{A}_{k+1} = Q_k \dots Q_1 (A + H_k) \quad (15.39)$$

и, согласно (15.38),

$$\|H_k\|_F \leq (6m-3k+40)k\eta + O(\eta^2). \quad (15.40)$$

## АНАЛИЗ ПОГРЕШНОСТЕЙ ОКРУГЛЕНИЙ ДЛЯ ЗАДАЧИ НК

В этой главе анализ погрешностей округления, проведенный в гл. 15, будет применен к выводу оценок для вычислительных погрешностей алгоритмов из гл. 11, 13, 14. Эти алгоритмы осуществляют решение шести типов задачи НК, изображенных на рис. 1.1.

Мы предполагаем в этой главе, что вся арифметика выполняется с одной и той же точностью  $\eta$ . Вследствие этого оценки погрешностей будут содержать множители  $n^2$  и  $mn$ . Можно получить меньшие оценки, зависящие лишь от первой степени  $n$  и вовсе не зависящие от  $m$ , если использовать арифметику повышенной точности в некоторых критических "узлах" численного процесса. Теоремы этой главы будут переформулированы в гл. 17, для арифметики со смешанной точностью.

**Теорема 16.1** (переопределенная задача наименьших квадратов полного ранга). Пусть  $A - m \times n$ -матрица псевдоранга  $n$ ,  $b - n$ -вектор. Если машинная арифметика имеет относительную точность  $\eta$ ,  $\bar{x}$  — решение задачи  $Ax \cong b$ , вычисленное посредством алгоритмов HFT (см. 11.4) и HS1 (см. 11.10) или HFTI (см. 14.9), то  $\bar{x}$  является точным решением задачи

$$(A + E)x \cong b + f, \quad (16.2)$$

где

$$\|E\|_F \leq (6m - 3n + 41)n\eta \|A\|_F + O(\eta^2), \quad (16.3)$$

$$\|f\| \leq (6m - 3n + 40)n\eta \|b\| + O(\eta^2). \quad (16.4)$$

**Доказательство.** Математически процесс можно описать формулами

$$\hat{Q}[A : b] = \begin{bmatrix} R & c \\ 0 & d \end{bmatrix}, \quad (16.5)$$

$$Rx = c. \quad (16.6)$$

Заменяя  $A$  в (15.39) и (15.40) матрицей  $[A : b]$  из (16.5), заключаем, что существуют ортогональная матрица  $Q$ , матрица  $G$  и вектор  $f$  такие, что вычисленные величины  $\bar{R}$ ,  $\bar{c}$  и  $\bar{d}$  удовлетворяют равенству

$$Q[A + G : b + f] = \begin{bmatrix} \bar{R} & \bar{c} \\ 0 & \bar{d} \end{bmatrix}, \quad (16.7)$$

где

$$\|G\|_F \leq (6m - 3n + 40)n\eta \|A\|_F + O(\eta^2), \quad (16.8)$$

$$\|f\| \leq (6m - 3n + 40)n\eta \|b\| + O(\eta^2). \quad (16.9)$$

Вместо (16.6), машина определяет  $x$  из системы  $\bar{R}x = \bar{c}$ . Согласно (15.45) и (15.46), вычисленное решение  $\bar{x}$  будет точно удовлетворять

Равенство (15.39) и оценку (15.40) можно интерпретировать следующим образом: вычисленная матрица  $\bar{A}_{k+1}$  есть точный результат ортогонального преобразования матрицы  $A + H_k$ , где  $\|H_k\|_F$  мала в сравнении с  $\|A\|_F$ .

Заметим, что результат, выражаемый соотношениями (15.39) и (15.40), не зависит от того факта, что матрицы  $\bar{Q}_i$  вычислялись с целью аннулировать некоторые элементы в  $\bar{A}_i$ . Эти соотношения останутся верными и в тех случаях, когда  $A$  (следовательно, и  $\bar{A}_{k+1}$ ) заменяется матрицей или вектором, не имеющими никакой специальной связи с матрицами  $\bar{Q}_i$ .

Далее, если матрицы, вычисленные процессом (15.34), используют в обратном порядке и

$$\bar{Y} = \text{fl}(\bar{Q}_1 \dots \bar{Q}_k X), \quad (15.41)$$

где  $X$  — произвольная матрица, то

$$\bar{Y} = Q_1 \dots Q_k (X + K), \quad (15.42)$$

причем

$$\|K\|_F \leq (6m - 3k + 40)k\eta \|X\|_F + O(\eta^2). \quad (15.43)$$

Ортогональные матрицы  $Q_i$  в (15.42) те же, что и в (15.39). Это замечание понадобится нам в доказательствах теорем 16.18, 16.36, 17.19 и 17.23.

Наконец, нам потребуются оценки погрешностей округления при решении треугольной системы уравнений. Пусть  $R$  — треугольная  $k \times k$ -матрица а  $c$  —  $k$ -вектор. Тогда вычисленное решение  $\bar{x}$  системы

$$Rx = c \quad (15.44)$$

будет удовлетворять равенству

$$(R + S)\bar{x} = c, \quad (15.45)$$

где  $S$  — треугольная матрица той же конфигурации, что и  $R$ ; при этом

$$\|S\|_F \leq k \|R\|_F \eta + O(\eta^2). \quad (15.46)$$

Доказательство этого результата дано в [11]. В процессе вывода (15.46) в указанной книге показано заодно, что

$$|s_{ii}| \leq 2\eta |r_{ii}| + O(\eta^2), \quad i = 1, \dots, k, \quad (15.47)$$

поэтому для разумных значений  $\eta$  невырожденность  $R$  обеспечивает невырожденность  $R + S$ .

#### Упражнение

15.48. Пусть  $G$  — матрица Гивенса, вычисленная алгоритмом G1 (см. 10.25) в точной арифметике, а  $\bar{G}$  — соответствующая матрица, вычисленная в арифметике с относительной точностью  $\eta$ . Пусть  $z$  — произвольный машинно представимый 2-вектор. Вывести оценку вида

$$\|\text{fl}(\bar{G}z) - Gz\| \leq c\eta \|z\|$$

и определить значение константы  $c$ .

равенству

$$(\bar{R} + S) \bar{x} = \bar{c},$$

причем

$$\|S\|_F \leq n\eta \|\bar{R}\|_F + O(\eta^2) = n\eta \|A\|_F + O(\eta^2). \quad (16.10)$$

Чтобы связать вычисленный вектор  $\bar{x}$  с исходной задачей наименьших квадратов, определим окаймленную матрицу

$$\begin{bmatrix} \bar{R} + S & \bar{c} \\ 0 & \bar{d} \end{bmatrix}.$$

Умножая ее слева на матрицу  $Q^T$  ( $Q$  — ортогональная матрица из (16.7)) получим

$$\begin{bmatrix} A + G + Q^T \begin{bmatrix} S \\ 0 \end{bmatrix} : b + f \end{bmatrix}.$$

Положим

$$E = G + Q^T \begin{bmatrix} S \\ 0 \end{bmatrix}.$$

Тогда  $\bar{x}$  будет решением задачи (16.2). Из неравенства  $\|E\| \leq \|G\| + \|S\|$  и формул (16.8), (16.10) вытекают оценки (16.3) и (16.4). Это завершает доказательство теоремы 16.1.

**Теорема 16.11 (квадратная невырожденная задача).** Пусть  $A$  —  $n \times n$ -матрица псевдоранга  $n$ ,  $a$   $b$  — некоторый  $n$ -вектор. Если машинная арифметика имеет относительную точность  $\eta$ , а  $\bar{x}$  — решение задачи  $Ax = b$ , вычисленное посредством алгоритмов HFT (см. 11.4) и HS1 (см. 11.10) или HFTI (см. 14.9), то  $\bar{x}$  является точным решением задачи

$$(A + E) \bar{x} = b + f, \quad (16.12)$$

где

$$\|E\|_F \leq (3n^2 + 41n) \eta \|A\|_F + O(\eta^2), \quad (16.13)$$

$$\|f\| \leq (3n^2 + 40n) \eta \|b\| + O(\eta^2). \quad (16.14)$$

Эта теорема — просто специальный случай теоремы 16.1, так что оценки (16.13) и (16.14) получаются при  $n = m$  из оценок (16.3) и (16.4). Мы сформулировали теорему 16.11 как отдельное утверждение, поскольку квадратный невырожденный случай часто представляет самостоятельный интерес.

Почувствительно сравнить теорему 16.11 с аналогичной теоремой о погрешностях для гауссова исключения с выбором главного элемента [11]. В последнем случае выполняется неравенство

$$\|E\|_\infty \leq 1,01(n^3 + 3n^2) \|A\|_\infty \eta \rho_n. \quad (16.15)$$

Здесь символ  $\|\cdot\|_\infty$  обозначает норму (максимальную строчную сумму), определяемую для произвольной  $m \times n$ -матрицы  $B$  формулой

$$\|B\|_\infty = \max \left\{ \sum_{j=1}^n |b_{ij}| : 1 \leq i \leq m \right\}.$$

Величина  $\rho_n$  является мерой относительного роста элементов в ходе гауссова исключения [11]. Если производится частичный выбор главного элемента, то имеет место оценка

$$\rho_n \leq 2^{n-1}. \quad (16.16)$$

Если используется полный выбор главного элемента, то

$$\rho_n \leq 1,8n^{(\log n)/4}. \quad (16.17)$$

Обычные реализации гауссова исключения опираются на стратегию частичного выбора. Поэтому оценка (16.15) для нормы матрицы  $E$  растет экспоненциально с ростом  $n$ .

При использовании алгоритмов Хаусхолдера, например алгоритма HFTI (см. 14.9), ситуация значительно более удовлетворительная и надежная. В этом случае в оценке (16.13) вообще нет коэффициента роста  $\rho_n$ . В [7] также отмечается этот факт, но указывается на то, что в алгоритме Хаусхолдера требуется примерно вдвое больше операций, чем в гауссовом исключении с частичным выбором. Кроме того, величину  $\rho_n$  легко вычислить в гауссовом алгоритме, и, согласно утверждению Уилкинсона, на практике она редко превышает 4, 8 или 16. Аналогичные замечания справедливы в отношении теоремы 17.15, где в некоторых узлах алгоритма Хаусхолдера используется повышенная точность.

В теоремах 16.1 и 16.11 было показано, что вычисленное решение является точным решением возмущенной задачи. В отличие от этого теоремы 16.18 и 16.36 утверждают, что вычисленное решение близко к нормальному решению возмущенной задачи. Эта более сложная формулировка связана с тем, что теоремы 16.18 и 16.36 относятся к задачам, решения которых были бы неединственными без требования минимальности длины.

**Теорема 16.18 (нормальное решение недоопределенной задачи полного ранга).** Пусть  $A$  —  $m \times n$ -матрица псевдоранга  $m$ ,  $a$   $b$  — некоторый  $m$ -вектор. Если машинная арифметика имеет относительную точность  $\eta$ , а  $\bar{x}$  — решение задачи  $Ax = b$ , вычисленное посредством алгоритмов HBT (см. 13.8) и HS2 (см. 13.9), то  $\bar{x}$  близко к точному решению возмущенной задачи в следующем смысле: существуют матрица  $E$  и вектор  $\hat{x}$  такие, что  $\hat{x}$  есть нормальное решение задачи

$$(A + E) \hat{x} = b, \quad (16.19)$$

причем

$$\|E\|_F \leq (6n - 3m + 41) m \eta \|A\|_F + O(\eta^2), \quad (16.20)$$

$$\|\bar{x} - \hat{x}\| \leq (6n - 3m + 40) m \eta \|\bar{x}\| + O(\eta^2). \quad (16.21)$$

**Доказательство.** Краткая математическая формулировка алгоритма гл. 13 такова:

$$A \hat{Q} = [R : 0], \quad (16.22)$$

$$Ry = b, \quad (16.23)$$

$$x = \hat{Q} \begin{bmatrix} y \\ 0 \end{bmatrix}. \quad (16.24)$$



Вместо (16.22) вычисленная нижняя треугольная матрица  $\bar{R}$  будет удовлетворять соотношению

$$(A + G) Q = [\bar{R} : 0], \quad (16.25)$$

где  $Q$  ортогональная и, согласно (15.40),

$$\|G\|_F \leq (6n - 3m + 40) m \eta \|A\|_F + O(\eta^2). \quad (16.26)$$

Далее из (15.45) и (15.46) следует, что вместо (16.23) вычисленный вектор  $\bar{y}$  будет точным решением системы

$$(\bar{R} + S) \bar{y} = b, \quad (16.27)$$

причем

$$\|S\| \leq m \eta \|\bar{R}\|_F + O(\eta^2) = m \eta \|A\|_F + O(\eta^2). \quad (16.28)$$

Вместо (16.24) вычисленный вектор  $\bar{x}$  определяется формулой

$$\bar{x} = Q \left( \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix} + h \right), \quad (16.29)$$

в которой, согласно (15.41)–(15.43),

$$\|h\| \leq (6n - 3m + 40) m \eta \|\bar{y}\| + O(\eta^2). \quad (16.30)$$

Ортогональная матрица  $Q$  в (16.29) та же, что и в (16.25), как указано в замечаниях, сопровождающих формулы (15.41)–(15.43).

Перепишывая (16.27) в виде

$$[\bar{R} + S : 0] \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix} = b, \quad (16.31)$$

затем в виде

$$[\bar{R} + S : 0] Q^T Q \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix} = b \quad (16.32)$$

и, наконец, используя (16.25) и (16.29), получаем

$$(A + G + [S : 0] Q^T) (\bar{x} - Qh) = b. \quad (16.33)$$

Положим

$$E = G + [S : 0] Q^T, \quad (16.34)$$

$$\hat{x} = \bar{x} - Qh \equiv Q \begin{bmatrix} \bar{y} \\ 0 \end{bmatrix}. \quad (16.35)$$

Исходя из неравенств (15.47), можно предполагать матрицу  $\bar{R} + S$  в (16.32) невырожденной. Поэтому пространство строк матрицы  $A + E \equiv [\bar{R} + S : 0] Q^T$  совпадает с оболочкой первых  $m$  столбцов матрицы  $Q$ .

Согласно (16.33)–(16.35),  $\hat{x}$  удовлетворяет уравнению (16.19); в то же время (16.35) показывает, что  $\hat{x}$  принадлежит оболочке первых  $m$  столбцов  $Q$ . Тем самым  $\hat{x}$  является единственным нормальным решением (16.19).

Для матрицы  $E$  и вектора  $\bar{x} - \hat{x}$  справедливы оценки

$$\|E\|_F \leq \|G\|_F + \|S\|_F \leq$$

$$\leq (6n - 3m + 41) m \eta \|A\|_F + O(\eta^2),$$

$$\|\bar{x} - \hat{x}\| = \|h\| \leq (6n - 3m + 40) m \eta \|\bar{y}\| + O(\eta^2) =$$

$$= (6n - 3m + 40) m \eta \|\bar{x}\| + O(\eta^2),$$

которые совпадают соответственно с (16.20) и (16.21). Теорема 16.18 доказана.

**Теорема 16.36 (задача НК неполного ранга).** Пусть  $A$  —  $m \times n$ -матрица,  $b$  —  $m$ -вектор. Пусть машинная арифметика имеет относительную точность  $\eta$  и  $\bar{x}$  — решение задачи  $Ax \cong b$ , вычисленное посредством алгоритма HFTI (см. 14.9). Пусть  $k$  — значение псевдоранга, определенное алгоритмом, а  $\bar{R}_{22}$  — вычисленная матрица, соответствующая матрице  $R_{22}$  из (14.1). Тогда  $\bar{x}$  близко к точному решению возмущенной задачи в следующем смысле: существуют матрица  $E$  и векторы  $f$  и  $\hat{x}$  такие, что  $\hat{x}$  является нормальным псевдорешением задачи

$$(A + E) x \cong b + f, \quad (16.37)$$

причем

$$\|E\|_F \leq \|\bar{R}_{22}\|_F + (6m + 6n - 6k - 3\mu + 84) \mu \eta \|A\|_F + O(\eta^2),$$

$$\mu = \min(m, n), \quad (16.38)$$

$$\|f\| \leq (6m - 3k + 40) k \eta \|b\| + O(\eta^2), \quad (16.39)$$

$$\|\bar{x} - \hat{x}\| \leq (6n - 6k + 43) k \eta \|\bar{x}\| + O(\eta^2). \quad (16.40)$$

**Доказательство.** Математическое описание алгоритма дано формулами (14.1)–(14.6). Напомним, что матрица  $Q$  в (14.1) является произведением  $n$  преобразований Хаусхолдера:  $Q = Q_n \dots Q_1$ . Заметим, что матрицы  $R_{11}$  и  $R_{12}$  из (14.1) и вектор  $c_1$  из (14.2) полностью определяются первыми  $k$  из этих преобразований. Пусть

$$\tilde{Q} = Q_k \dots Q_1. \quad (16.41)$$

Тогда

$$\tilde{Q} A P = \begin{bmatrix} R_{11} & R_{12} \\ 0 & S \end{bmatrix}, \quad (16.42)$$

$$\tilde{Q} b = \begin{bmatrix} c_1 \\ d \end{bmatrix}, \quad (16.43)$$

причем

$$\|S\|_F = \|\bar{R}_{22}\|_F, \quad (16.44)$$

$$\|d\| = \|c_2\|. \quad (16.45)$$

Вместо (16.42) вычисленные матрицы  $\bar{R}_{11}, \bar{R}_{12}, \bar{S}$  удовлетворяют соотношению

$$\hat{Q}(A+G)P = \begin{bmatrix} \bar{R}_{11} & \bar{R}_{12} \\ 0 & \bar{S} \end{bmatrix}. \quad (16.46)$$

Согласно результатам гл. 15,  $\hat{Q}$  — точно ортогональная матрица и

$$\|G\|_F \leq \left( \sum_{l=1}^k \alpha_{m+1-l} \right) \|A\|_F + O(\eta^2), \quad (16.47)$$

$$\|\bar{S}\|_F \leq \|\bar{R}_{22}\|_F + \left( \sum_{l=k+1}^{\mu} \alpha_{m+1-l} \right) \|A\|_F + O(\eta^2).$$

Величина  $\alpha_j$  определяется формулой (15.37);  $\mu = \min(m, n)$ , как и в гл. 14, а  $\bar{R}_{22}$  обозначает вычисленную матрицу, соответствующую  $R_{22}$  в (14.1).

Из (16.46) можно вывести формулу

$$\hat{Q}(A+H)P = \begin{bmatrix} \bar{R}_{11} & \bar{R}_{12} \\ 0 & 0 \end{bmatrix}, \quad (16.48)$$

где

$$\begin{aligned} \|H\|_F &\leq \|\bar{S}\|_F + \|G\|_F \leq \|\bar{R}_{22}\|_F + \left( \sum_{l=1}^{\mu} \alpha_{m+1-l} \right) \|A\|_F + O(\eta^2) \leq \\ &\leq \|\bar{R}_{22}\|_F + (6m - 3\mu + 40)\mu\eta \|A\|_F + O(\eta^2). \end{aligned} \quad (16.49)$$

Вместо (14.3) вычисленная матрица  $\bar{W}$  удовлетворяет равенству

$$\{[\bar{R}_{11} : \bar{R}_{12}] + M\} \hat{K} = [\bar{W} : 0], \quad (16.50)$$

где  $\hat{K}$  — ортогональная матрица.

Оценка для  $\|M\|_F$  несколько отличается от оценки (15.40). Дело в том, что каждое из  $k$  преобразований Хаусхолдера, чье произведение образует матрицу  $K$ , отличается от единичной матрицы только  $n - k + 1$  столбцами. Поэтому

$$\begin{aligned} \|M\|_F &\leq \alpha_{n-k+1} k\eta \|\bar{R}_{11} : \bar{R}_{12}\|_F + O(\eta^2) = \\ &= (6n - 6k + 43)k\eta \|A\|_F + O(\eta^2), \end{aligned} \quad (16.51)$$

где  $\alpha_j$  определены формулой (15.37).

Вместо (16.43) вычисленные векторы  $\bar{c}_1$  и  $\bar{d}$  удовлетворяют соотношению

$$\hat{Q}(b+f) = \begin{bmatrix} \bar{c}_1 \\ \bar{d} \end{bmatrix}, \quad (16.52)$$

где  $\hat{Q}$  — та же матрица, что и в (16.46), а

$$\|f\| \leq (6m - 3k + 40)k\eta \|b\| + O(\eta^2); \quad (16.53)$$

это и есть оценка (16.39) теоремы 16.36.

Что касается (14.4), то вычисленный вектор  $\bar{y}_1$  будет решением системы

$$(\bar{W} + Z) \bar{y}_1 = \bar{c}_1 \quad (16.54)$$

с невырожденной матрицей  $\bar{W} + Z$ ; при этом

$$\|Z\|_F \leq k\eta \|\bar{W}\|_F + O(\eta^2) \leq k\eta \|A\|_F + O(\eta^2). \quad (16.55)$$

В (14.5) мы рассмотрим только случай  $y_2 = 0$ . Вместо (14.6) вычисленный вектор  $\bar{x}$  удовлетворяет равенству

$$\bar{x} = P \hat{K} \begin{bmatrix} \bar{y}_1 \\ 0 \end{bmatrix} + h, \quad (16.56)$$

где

$$\|h\| \leq (6n - 6k + 43)k\eta \|\bar{y}_1\| + O(\eta^2). \quad (16.57)$$

Оценка (16.57) выводится таким же образом, как и (16.51). Положим

$$\hat{x} = P \hat{K} \begin{bmatrix} \bar{y}_1 \\ 0 \end{bmatrix}. \quad (16.58)$$

Тогда из (16.56) — (16.58) следует

$$\|\bar{x} - \hat{x}\| \leq (6n - 6k + 43)k\eta \|\bar{x}\| + O(\eta^2), \quad (16.59)$$

т.е. оценка (16.40).

Аналогично доказательству теоремы 16.18 из (16.54), (16.58) выводится, что  $\hat{x}$  есть нормальное решение задачи

$$[\bar{W} + Z : 0] \hat{K}^T P^T x = \bar{c}_1. \quad (16.60)$$

Согласно (16.50)

$$[\bar{W} + Z : 0] \hat{K}^T = [\bar{R}_{11} : \bar{R}_{12}] + X, \quad (16.61)$$

где

$$X = M + [Z : 0] \hat{K}^T; \quad (16.62)$$

при этом из (16.51) и (16.55) следует

$$\|X\|_F \leq \|M\|_F + \|Z\|_F \leq (6n - 6k + 44)k\eta \|A\|_F + O(\eta^2). \quad (16.63)$$

Подставляя (16.61) в (16.60) и окаймляя систему дополнительными строками, видим, что  $\hat{x}$  является нормальным псевдорешением задачи

$$\left( \begin{bmatrix} \bar{R}_{11} & \bar{R}_{12} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} X \\ 0 \end{bmatrix} \right) P^T x \cong \begin{bmatrix} \bar{c}_1 \\ \bar{d} \end{bmatrix}. \quad (16.64)$$

Умножая (16.64) слева на  $\hat{Q}^T$  и используя (16.46), заключаем, что  $\hat{x}$  есть нормальное псевдорешение задачи (16.37), где

$$E = H + \hat{Q}^T \begin{bmatrix} X \\ 0 \end{bmatrix} P^T. \quad (16.65)$$

Следовательно,

$$\|E\|_F \leq \|H\|_F + \|X\|_F. \quad (16.66)$$

Из (16.49) и (16.63), где  $k$  для простоты заменено числом  $\mu$  ( $\mu \geq k$ ), получаем окончательное выражение

$$\|E\|_F \leq \|\bar{R}_{22}\|_F + (6m + 6n - 6k - 3\mu + 84)\mu\eta \|A\|_F + O(\eta^2). \quad (16.67)$$

Вместе с (16.53) и (16.59) оно доказывает теорему 16.36.

## ГЛАВА 17

### АНАЛИЗ ПОГРЕШНОСТЕЙ ОКРУГЛЕНИЙ ДЛЯ ЗАДАЧИ НК В АРИФМЕТИКЕ СО СМЕШАННОЙ ТОЧНОСТЬЮ

В анализе погрешностей округлений мы предполагали до сих пор, что вся арифметика выполняется с одной и той же точностью, характеризуемой параметром  $\eta$ . Если в данной вычислительной машине эффективно реализована арифметика повышенной точности (характеризуемая параметром  $\omega$ ,  $\omega < \eta$ ) и имеются эффективные машинные команды для перевода чисел из одной разрядности в другую, то можно использовать эту повышенную точность в отдельных узлах численного процесса. Это позволяет улучшить точность результатов с очень малыми издержками по времени и, по существу, без дополнительных требований к памяти.

Если не решено перейти полностью с  $\eta$ -точности на  $\omega$ -точность, то обычно использование арифметики с  $\omega$ -точностью ограничивают теми частями алгоритма, где требуется лишь небольшое фиксированное (т.е. не зависящее от размерностей задачи  $m$  и  $n$ ) число ячеек с  $\omega$ -разрядностью.

Уилкинсон провел подробный анализ погрешностей округлений для алгоритмов линейной алгебры в случае конкретного варианта арифметики со смешанной точностью, а именно  $\omega = \eta^2$  (см. [7]). Мы воспользуемся сходными методами, чтобы вывести для алгоритмов, уже проанализированных в предыдущей главе, оценки в той же специальной арифметике со смешанной точностью.

При вычислении преобразования Хаусхолдера мы модифицируем процесс (15.13)–(15.19) следующим образом. Вычисления по формуле (15.14) будут выполняться в арифметике с  $\omega$ -точностью, а конечный результат будет представлен так, что  $\bar{T}$  – число  $\eta$ -разрядности. Выражения для погрешностей в (15.20)–(15.25) примут вид

$$\bar{T} = t(1 + \tau), \quad |\tau| \leq \eta + O(\eta^2), \quad (17.1)$$

$$\bar{s} = s(1 + \sigma), \quad |\sigma| \leq 5\eta/2 + O(\eta^2), \quad (17.2)$$

$$\bar{u}_1 = u_1(1 + \kappa), \quad |\kappa| \leq 7\eta/2 + O(\eta^2), \quad (17.3)$$

$$\bar{b} = b(1 + \beta), \quad |\beta| \leq 7\eta + O(\eta^2), \quad (17.4)$$

$$\|\bar{Q} - Q\|_F \leq 28\eta + O(\eta^2). \quad (17.5)$$

Вместо (15.27) матрица

$$\bar{B} = \Pi[C + \bar{u}(\bar{u}^T C)\bar{b}^{-1}] \quad (17.6)$$

будет вычислена в арифметике с  $\omega$ -точностью, а затем ее элементы будут переведены в форму с  $\eta$ -точностью. Поэтому оценки погрешностей (15.29) и (15.30) заменятся соответственно на

$$\|E\|_F \leq \|C\|_F \eta + O(\eta^2) \quad (17.7)$$

и (с учетом (17.5)) на

$$\|\bar{B} - B\|_F \leq 29\|C\|_F \eta + O(\eta^2). \quad (17.8)$$

Теперь оценка (15.40) для погрешности, соответствующей умножению на  $k$  преобразований Хаусхолдера, превращается в

$$\|H_k\|_F \leq 29k\|A\|_F \eta + O(\eta^2). \quad (17.9)$$

Наконец, при решении треугольной системы (15.44) тоже можно было бы использовать арифметику с  $\omega$ -точностью. В этом случае вместо оценки (15.46) мы имели бы

$$\|S\|_F \leq \|R\|_F \eta + O(\eta^2). \quad (17.10)$$

Оценка (15.46) уже и так мала по сравнению с оценками погрешностей для других этапов полного решения задачи НК. Следовательно, при использовании  $\omega$ -точности в решении (15.44) уменьшение оценки для общей погрешности процесса будет очень незначительно. По этой причине мы предполагаем, что система (15.44) решается в  $\eta$ -точности, а не в  $\omega$ -точности.

Для версий наших алгоритмов, использующих смешанную арифметику, теоремы 16.1, 16.11, 16.18 и 16.36 заменяются соответственно теоремами 17.11, 17.15, 17.19 и 17.23. Доказательства опущены, поскольку они мало чем отличаются от соответствующих доказательств гл. 16.

**Теорема 17.11 (переопределенная задача наименьших квадратов полного ранга).** Пусть  $A$  –  $m \times n$ -матрица псевдоранга  $n$ , а  $b$  – некоторый  $m$ -вектор. Пусть используется арифметика со смешанной точностью, характеризуемая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $\bar{x}$  – решение задачи  $Ax \cong b$ , вычисленное посредством алгоритмов HFT (см. 11.4) и HS1 (см. 11.10) или HFTI (см. 14.9). Тогда  $\bar{x}$  является точным решением задачи

$$(A + E)x \cong b + f, \quad (17.12)$$

где

$$\|E\|_F \leq 30n\|A\|_F \eta + O(\eta^2), \quad (17.13)$$

$$\|f\| \leq 29n\|b\| \eta + O(\eta^2). \quad (17.14)$$

**Теорема 17.15 (квадратная невырожденная задача).** Пусть  $A$  –  $n \times n$ -матрица псевдоранга  $n$ , а  $b$  – некоторый  $n$ -вектор. Пусть используется арифметика со смешанной точностью, характеризуемая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $\bar{x}$  – решение задачи  $Ax = b$ , вычисленное посредством алгоритмов HFT (см. 11.4) и HS1 (см. 11.10) или HFTI (см. 14.9). Тогда  $\bar{x}$  является точным решением задачи

$$(A + E)x = b + f, \quad (17.16)$$

$$\|E\|_F \leq 30n \|A\|_F \eta + O(\eta^2), \quad (17.17)$$

$$\|f\| \leq 29n \|b\| \eta + O(\eta^2). \quad (17.18)$$

Теорема 17.19 (нормальное решение недоопределенной задачи полного ранга). Пусть  $A - m \times n$ -матрица псевдоранга  $m$ , а  $b$  — некоторый  $m$ -вектор. Пусть используется арифметика со смешанной точностью, характеризующая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $\bar{x}$  — решение задачи  $Ax = b$ , вычисленное посредством алгоритмов НВТ (см. 13.8) и HS2 (см. 13.9). Тогда  $\bar{x}$  близко к точному решению возмущенной задачи в следующем смысле: существуют матрица  $E$  и вектор  $\hat{x}$  такие, что  $\hat{x}$  является нормальным решением задачи

$$(A + E)x = b, \quad (17.20)$$

причем

$$\|E\|_F \leq 30m \|A\|_F \eta + O(\eta^2), \quad (17.21)$$

$$\|\bar{x} - \hat{x}\| \leq 29m \|\bar{x}\| \eta + O(\eta^2). \quad (17.22)$$

Теорема 17.23 (задача НК неполного ранга). Пусть  $A - m \times n$ -матрица,  $b - m$ -вектор. Пусть используется арифметика со смешанной точностью, характеризующая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $\bar{x}$  — решение задачи  $Ax \cong b$ , вычисленное посредством алгоритма HFTI (см. 14.9). Пусть  $k$  — значение псевдоранга, определенное алгоритмом, а  $\bar{R}_{22}$  — вычисленная матрица, соответствующая матрице  $R_{22}$  из (14.1). Тогда  $\bar{x}$  близко к точному решению возмущенной задачи в следующем смысле: существуют матрица  $E$  и векторы  $f$  и  $\hat{x}$  такие, что  $\hat{x}$  является нормальным псевдорешением задачи

$$(A + E)x \cong b + f, \quad (17.24)$$

причем

$$\|E\|_F \leq \|\bar{R}_{22}\|_F + 59k \|A\|_F + O(\eta^2), \quad (17.25)$$

$$\|f\| \leq 29k \|b\| \eta + O(\eta^2), \quad (17.26)$$

$$\|\bar{x} - \hat{x}\| \leq 29k \|\bar{x}\| \eta + O(\eta^2). \quad (17.27)$$

Другого типа анализ погрешностей округлений для того же алгоритма — последовательности преобразований Хаусхолдера в применении к задаче  $A_{m \times n} x \cong b$ ,  $m \geq n$ ,  $\text{rank } A = n$  — был выполнен в работах [146, 147]. Этот анализ был мотивирован рассмотрением задач наименьших квадратов с сильно различающимися весами. Такие задачи являются средством решения задачи НК с ограничениями равенствами и в этом качестве исследуются в гл. 22.

В гл. 22 мы рассмотрим задачу наименьших квадратов с таким свойством: все элементы некоторых строк исходной матрицы  $[A : b]$  очень малы по величине сравнительно с элементами других строк; тем не менее относительная точность этих малых элементов существенна для задачи. Последнее означает, что возмущения этих элементов, значительные по отношению к их величинам, приводят к большим изменениям решения.

Ясно, что теорема 17.11 не гарантирует, что алгоритм Хаусхолдера даст удовлетворительное решение такой задачи; ведь неравенства (17.13) и (17.14) допускают возмущения всех элементов, которые, хотя и "малы" сравнительно с наибольшими элементами в  $A$  и  $b$  соответственно, могут быть велики по отношению к малым элементам.

И действительно, Пауэлл и Рид обнаружили экспериментально, что результаты, получаемые алгоритмом Хаусхолдера (таким например, как алгоритм HFTI (см. 14.9)) для задач с сильно различающимися весами, критически зависят от упорядочения строк матрицы  $[A : b]$ . Они заметили, что если в качестве ведущей берется строка из малых элементов, а некоторые последующие строки имеют большие элементы в ведущем столбце, то в (10.7) величина  $v_p$  будет мала сравнительно с  $s$ . В предельном случае, когда  $|v_p|/|s| < \eta$ , величина  $v_p$  вовсе не дает вклада в  $u_p$ . Кроме того, в подобных случаях  $p$ -я компонента вектора  $c_j$  в (10.21) может оказаться очень малой относительно  $p$ -й компоненты вектора  $t_{ji}$  и не даст вклада в вектор  $\tilde{c}_j$ . Таким образом, если элементы ведущей строки очень малы в сравнении с элементами некоторых последующих строк, то эффект, по существу, будет тот же, как если бы мы заменили коэффициенты ведущей строки нулями. Если решение чувствительно к подобной потере информации, то описанной ситуации следует по возможности избегать.

Пауэлл и Рид получили удовлетворительные результаты для таких задач, когда дополнили алгоритм стратегией перестановок строк; она будет описана перед формулировкой теоремы 17.37.

Ниже будут сформулированы три теоремы [146], которые позволяют понять поведение алгоритма Хаусхолдера для задач с сильно различающимися весами.

Будет удобно считать, что столбцы  $A$  упорядочены заранее, так что при выполнении алгоритма HFTI перестановок столбцов не происходит. Положим  $[\bar{A}^{(1)} : \bar{b}^{(1)}] = [A : b]$ , и пусть  $[\bar{A}^{(i+1)} : \bar{b}^{(i+1)}]$  обозначает вычисленный результат применения  $i$ -го преобразования Хаусхолдера к матрице  $[\bar{A}^{(i)} : \bar{b}^{(i)}]$ ,  $i = 1, \dots, n$ .

Введем обозначение

$$\gamma_i = \max_{j,k} |\bar{a}_{ij}^{(k)}|, \quad i = 1, \dots, m, \quad (17.28)$$

где  $\bar{a}_{ij}^{(k)}$  — элемент  $(i, j)$  матрицы  $\bar{A}^{(k)}$ .

Пусть  $Q^{(k)}$  — точная ортогональная матрица Хаусхолдера, определяемая условием, чтобы элементы в позициях  $(i, k)$ ,  $i = k + 1, \dots, m$ , матричного произведения  $Q^{(k)} \bar{A}^{(k)}$  были нулями.

Теорема 17.29 [146]. Пусть  $A - m \times n$ -матрица псевдоранга  $n$ . Пусть используется арифметика со смешанной точностью, характеризующая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $A$  приведена к верхней треугольной  $m \times n$ -матрице  $\bar{A}^{(n+1)}$  посредством алгоритма HFTI (см. 14.9). Тогда  $\bar{A}^{(n+1)} = Q^{(n)} \dots Q^{(1)} (A + E)$ , (17.30)

причем элементы  $e_{ij}$  матрицы  $E$  удовлетворяют оценкам

$$|e_{ij}| \leq (20n^2 - 17n + 44) \gamma_i \eta + O(\eta^2). \quad (17.31)$$

Чтобы сформулировать теорему 17.32, нам потребуются дополнительные определения, учитывающие то обстоятельство, что правая часть  $b$  не может участвовать в перестановках столбцов:

$$\mu_i = \max_k |\bar{b}_i^{(k)}|, \quad i = 1, \dots, m,$$

$$\nu_k = \left( \sum_{i=k}^m [\bar{b}_i^{(k)}]^2 \right)^{1/2}, \quad \sigma_k = \left( \sum_{i=k}^m [\bar{a}_{ik}^{(k)}]^2 \right)^{1/2}, \quad k = 1, \dots, n,$$

$$\rho = \max \left( \max_i \frac{\mu_i}{\gamma_i}, \max_k \frac{\nu_k}{\sigma_k} \right).$$

Отметим, что из наших предположений об **априорном упорядочении** столбцов  $A$  следует

$$\sigma_k = \max_{k < j \leq n} \left( \sum_{i=k}^m [\bar{a}_{ij}^{(k)}]^2 \right)^{1/2}, \quad k = 1, \dots, n.$$

**Теорема 17.32 [146].** Пусть  $A$  —  $m \times n$ -матрица псевдоранга  $n$ , а  $b$  — некоторый  $m$ -вектор. Пусть используется арифметика со смешанной точностью, характеризуемая параметрами  $\eta$  и  $\omega \leq \eta^2$ , и пусть  $\bar{x}$  — решение задачи  $Ax \cong b$ , вычисленное посредством алгоритма HFTI (см. 14.9). Тогда  $\bar{x}$  является решением задачи

$$(A + E)x \cong b + f, \quad (17.33)$$

причем выполняются следующие оценки для элементов  $e_{ij}$  матрицы  $E$  и элементов  $f_i$  вектора  $f$ :

$$|e_{ij}| \leq (20n^2 - 15n + 44) \gamma_i \eta + O(\eta^2), \quad (17.34)$$

$$|f_i| \leq (20n^2 + 20n + 44) \rho \gamma_i \eta + O(\eta^2). \quad (17.35)$$

Используя неравенство  $|e_{ij}| \leq \|E\|_F$ , из (17.13) можно вывести

$$|e_{ij}| \leq 30n \|A\|_F \eta + O(\eta^2). \quad (17.36)$$

Сравнивая неравенства (17.34) и (17.36), видим, что оценка (17.34) будет наиболее полезна в тех случаях, когда  $\gamma_i$  существенно меньше, чем  $\|A\|_F$ . Аналогичные замечания справедливы в отношении оценок (17.14) и (17.35).

Напомним, что  $\gamma_i$  обозначает величину наибольшего элемента в  $i$ -х строках всех матриц  $A^{(1)}, \dots, A^{(n+1)}$ . Поэтому, чтобы  $\gamma_i$  была мала (относительно  $\|A\|_F$ ), нужно, чтобы были малы элементы в  $i$ -й строке исходной матрицы  $A^{(1)} = A$  и чтобы элементы  $i$ -х строк последующих матриц  $A^{(k)}$  не возрастали сколько-нибудь значительно.

Пауэлл и Рид привели пример, показывающий, что если не накладывать ограничений на упорядочение строк  $A$ , то рост поначалу малых элементов в некоторых строках может быть очень большим. Они рекомендуют следующую стратегию перестановок строк. Предположим, что перед выполнением  $k$ -го преобразования Хаусхолдера ведущий столбец выбран и переведен в  $k$ -й столбец хранящего массива. Определим строчный индекс  $l$  такой, что

$$|a_{lk}| = \max \{ |a_{ik}| : k \leq i \leq m \}.$$

Переставим строки  $l$  и  $k$ . Теперь  $k$ -е преобразование Хаусхолдера строится и применяется обычным образом. Этот процесс выполняется для  $k = 1, \dots, n$ .

**Теорема 17.37 [146].** Если используется только что описанная стратегия перестановок столбцов и строк, то величины  $\gamma_i$ , определенные в (17.28), удовлетворяют оценке

$$\gamma_i \leq (1 + 2^{1/2})^{n-1} m^{1/2} \max_j |a_{ij}^{(1)}|$$

для  $1 \leq i \leq m$ .

Пауэлл и Рид сообщают, что, применяя эту комбинированную стратегию перестановок столбцов и строк, они получили удовлетворительные решения для некоторых задач с сильно различающимися весами. Для тех же задач результаты были неудовлетворительны, когда допускалось использование строк с малыми элементами в качестве ведущих строк.

## ГЛАВА 18

### ВЫЧИСЛЕНИЕ СИНГУЛЯРНОГО РАЗЛОЖЕНИЯ И РЕШЕНИЕ ЗАДАЧИ НК

#### § 1. Введение

Сингулярное разложение матрицы было темой основополагающей работы [73]. В этой работе содержалась библиография, относящаяся к приложениям и алгоритмам, и была начата разработка нового метода, законченная в основном публикацией [34]. В книге [8] помещена усовершенствованная версия этого метода. Она представляет собой специальную адаптацию  $QR$ -алгоритма [59] для вычисления собственных значений и собственных векторов симметричной матрицы.

В этой главе будут описаны  $QR$ -алгоритм для симметричных матриц и (несколько модифицированный) алгоритм Голуба и Райнша для вычисления сингулярного разложения

$$A_{m \times n} = U_{m \times m} \begin{bmatrix} S_{n \times n} \\ 0 \end{bmatrix} V_{n \times n}^T.$$

Будет разобран также вопрос об использовании сингулярного разложения для анализа и решения задачи НК.

Чтобы не менять обозначений, а также потому, что применение сингулярного анализа в этом случае наиболее вероятно, мы предполагаем на протяжении всей главы, что  $m \geq n$ . Случай  $m < n$  можно свести к предыдущему, приписывая  $n - m$  нулевых строк к матрице  $A$  или в случае задачи НК к расширенной матрице коэффициентов  $[A : b]$ . Машинная программа может выполнять окаймление нулями неявно, так что не потребуются реального хранения нулевых строк или арифметических операций с нулевыми элементами.

## § 2. QR-алгоритм для симметричных матриц

QR-алгоритм Фрэнсиса \*) для случая симметричных матриц и при условии использования сдвигов является одним из самых успешных методов во всем численном анализе. Имеется ясное представление о его свойствах и скорости сходимости. В этом параграфе мы опишем алгоритм и заложим основы для доказательства его сходимости.

Вначале симметричную  $n \times n$ -матрицу  $A$  можно преобразовать к трехдиагональному виду посредством ортогональных преобразований подобия:  $n - 2$  преобразований Хаусхолдера или  $(n - 1)(n - 2)/2$  преобразований Гивенса. Поэтому без потери общности мы можем считать, что симметричная матрица, для которой нужно вычислить собственные значения, трехдиагональная.

QR-алгоритм со сдвигами для вычисления спектрального разложения симметричной трехдиагональной  $n \times n$ -матрицы  $A$  можно описать формулами

$$A_1 = A, \quad (18.1)$$

$$Q_k(A_k - \sigma_k I_n) = R_k, \quad (18.2)$$

$$R_k Q_k^T + \sigma_k I_n = A_{k+1}, \quad (18.3)$$

где для  $k = 1, 2, \dots$

- а)  $Q_k$  — ортогональная матрица;
- б)  $R_k$  — верхняя треугольная матрица;
- с)  $\sigma_k$  —  $k$ -й сдвиг.

Прежде чем описать метод вычисления сдвигов  $\sigma_k$ , мы введем обозначения для элементов матриц  $A_k$ . Можно проверить (см. упражнение 18.46), что для трехдиагональной матрицы  $A_1$  все матрицы  $A_k$ ,  $k = 2, 3, \dots$ , также будут трехдиагональными. В этой главе и в приложении В мы будем обозначать диагональные элементы каждой трехдиагональной матрицы  $A_k$  через  $a_i^{(k)}$ ,  $i = 1, \dots, n$ , а наддиагональные и поддиагональные элементы через  $b_i^{(k)}$ ,  $i = 2, \dots, n$ . Теперь мы можем определить сдвиг  $\sigma_k$  следующим образом.

Каждый сдвиг  $\sigma_k$  есть то собственное значение нижней угловой  $2 \times 2$ -подматрицы матрицы  $A_k$

$$\begin{bmatrix} a_{n-1}^{(k)} & b_n^{(k)} \\ b_n^{(k)} & a_n^{(k)} \end{bmatrix}, \quad (18.4)$$

которое находится ближе к  $a_n^{(k)}$ .

Из (18.1)–(18.3) следует

$$A_{k+1} = Q_k A_k Q_k^T,$$

так что все матрицы  $A_k$  имеют одни и те же собственные значения. Обозначим их через  $\lambda_1, \dots, \lambda_n$ .

\*) QR-алгоритм был изобретен советским алгебраистом В.Н. Кублановской независимо от Фрэнсиса и одновременно с ним. (Примеч. пер.)

Если для некоторого  $k$  среди элементов  $b_i^{(k)}$  есть нулевые, то матрица  $A_k$  разлагается в прямую сумму подматриц, в каждой из которых все наддиагональные и поддиагональные элементы отличны от нуля. Поэтому нам достаточно рассмотреть задачу, в которой все  $b_i^{(k)}$  ненулевые.

Далее, согласно лемме В.1 приложения В, из того, что все поддиагональные элементы не равны нулю, вытекает, что все собственные значения простые. Таким образом, мы можем ограничиться вычислением спектрального разложения для симметричных трехдиагональных матриц с простыми собственными значениями.

Сходимость QR-алгоритма устанавливается следующей теоремой [194, 195].

**Теорема 18.5 (глобальная квадратичная сходимость QR-алгоритма со сдвигами).** Пусть  $A = A_1$  — симметричная трехдиагональная  $n \times n$ -матрица с ненулевыми поддиагональными элементами. Пусть  $A_k$  — матрицы, ортогонально подобные матрице  $A$ , которые генерируются QR-алгоритмом со сдвигами в соответствии с формулами (18.1)–(18.3) и (18.4). Тогда а) каждая матрица  $A_k$  трехдиагональная и симметричная;

б) элемент  $b_n^{(k)}$  матрицы  $A_k$ , стоящий на пересечении  $n$ -й строки и  $n - 1$ -го столбца, стремится к нулю при  $k \rightarrow \infty$ ;

с) сходимость  $b_n^{(k)}$  к нулю асимптотически квадратичная, т.е. существует  $\epsilon > 0$ , зависящее от  $A$  и такое, что для всех  $k$

$$|b_n^{(k+1)}| \leq \epsilon |b_n^{(k)}|^2.$$

Доказательство этой теоремы дано в приложении В.

На практике сходимость алгоритма обычно бывает кубической. Однако квадратичная сходимость — это лучшее, что удастся доказать. Другие замечания по этому поводу читатель найдет в приложении В.

## § 3. Вычисление сингулярного разложения

Рассмотрим теперь построение сингулярного разложения  $m \times n$ -матрицы  $A$  в предположении, что  $m \geq n$ . Замечания по поводу случая  $m < n$  см. в § 1.

Сингулярное разложение будет вычислено в два этапа. На первом этапе  $A$

преобразуется к верхней двухдиагональной матрице  $\begin{bmatrix} B \\ 0 \end{bmatrix}$  посредством последовательности (не более чем из  $2n - 1$ ) преобразований Хаусхолдера

$$\begin{bmatrix} B \\ 0 \end{bmatrix} = Q_n (\dots ((Q_1 A) H_2) \dots H_n) \equiv Q^T A H, \quad (18.6)$$

где

$$B = \begin{bmatrix} q_1 e_2 & & & \\ & q_2 e_3 & & \\ & & \ddots & \\ & & & q_{n-1} e_n \\ & & & & q_n \end{bmatrix} \quad (18.7)$$

трансформирующая матрица  $Q_i$  выбирается так, чтобы аннулировать элементы  $i+1, \dots, m$  столбца  $i$ ; матрица  $H_i$  — так, чтобы аннулировать элементы  $i+1, \dots, n$  строки  $i-1$ .

Заметим, что  $H_n$  — это попросту единичная матрица. Она включена в (18.6), чтобы упростить обозначения;  $Q_n$  также будет единичной матрицей при  $m = n$ , но при  $m > n$  она, вообще говоря, отличается от единичной.

Второй этап процесса состоит в применении специальным образом адаптированного  $QR$ -алгоритма к вычислению сингулярного разложения матрицы  $B$

$$B = \hat{U} S \hat{V}^T; \quad (18.8)$$

здесь  $\hat{U}$  и  $\hat{V}$  — ортогональные матрицы, а  $S$  диагональная. Из (18.8) можно получить сингулярное разложение  $A$ :

$$A = (Q\hat{U}) \begin{bmatrix} S \\ 0 \end{bmatrix} (H\hat{V})^T \equiv U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T. \quad (18.9)$$

Обсудим теперь вычисление разложения (18.8). Заметим прежде всего, что если какой-либо элемент  $e_i$  равен нулю, то матрица  $B$  распадается на блоки, которые можно обрабатывать независимо друг от друга. Ниже мы покажем, что если какой-либо элемент  $q_i$  равен нулю, то применение некоторых преобразований также позволяет привести матрицу  $B$  к распадающемуся виду.

Предположим, что  $q_k = 0$ , но  $q_j \neq 0$  и  $e_j \neq 0$  для  $j = k+1, \dots, n$ . Умножая  $B$  слева на  $n-k$  вращений Гивенса  $T_j$ , получим

$$T_n \dots T_{k+1} B = B' \equiv \begin{bmatrix} q_1 e_2 & & & \\ & q_{k-1} e_k & & \\ & 0 & 0 & \\ & & q'_{k+1} e'_{k+2} & \\ & & & \ddots \\ & & & q'_{n-1} e'_n \\ & & & & q'_n \end{bmatrix}, \quad (18.10)$$

причем  $e'_j \neq 0$ ,  $j = k+2, \dots, n$ , и  $q'_j \neq 0$ ,  $j = k+1, \dots, n$ .

Вращение  $T_j$  конструируется так, чтобы аннулировать элемент, стоящий на пересечении строки  $k$  и столбца  $j$ ,  $j = k+1, \dots, n$ ; оно применяется к строкам  $k$  и  $j$ ,  $j = k+1, \dots, n$ .

Матрица  $B'$  в (18.10) имеет вид

$$B' = \begin{bmatrix} B'_1 & 0 \\ 0 & B'_2 \end{bmatrix},$$

где в  $B'_2$  все диагональные и наддиагональные элементы отличны от нуля. Прежде чем продолжить обсуждение  $B'_2$ , заметим, что  $B'_1$  имеет хотя бы одно нулевое сингулярное число, поскольку ее нижний угловой элемент (в позиции  $(k, k)$ ) равен нулю.

Когда позднее алгоритм возвращается к обработке матрицы  $B'_1$ , этот факт можно использовать, чтобы исключить  $e_k$  посредством следующей последовательности вращений:  $B'_1 = B'_1 R_{k-1} \dots R_1$ . Здесь  $R_i$  оперирует со столбцами  $i$  и  $k$  и аннулирует элемент в позиции  $(i, k)$ . При  $i > 1$  такое вращение создаст ненулевой элемент в позиции  $(i-1, k)$ .

Вернемся к рассмотрению матрицы  $B'_2$ . Будет удобно использовать прежний символ  $B$  для обозначения этой двухдиагональной матрицы с ненулевыми диагональными и наддиагональными элементами. Символ  $n$ , как и раньше, обозначает порядок  $B$ .

Сингулярное разложение (18.8) матрицы  $B$  будет получено посредством следующего итерационного процесса:  $B_1 = B$ ,  $B_{k+1} = U_k^T B_k V_k$ ,  $k = 1, 2, \dots$ . Здесь  $U_k$  и  $V_k$  — ортогональные матрицы, а  $B_k$  — верхняя двухдиагональная матрица для всех  $k$ . Матрицы  $U_k$  и  $V_k$  выбираются таким образом, что существует и диагональная матрица  $\tilde{S} = \lim_{k \rightarrow \infty} B_k$ .

Заметим, что диагональные элементы матрицы  $\tilde{S}$ , полученной непосредственно из этой итерационной процедуры, не являются в общем случае ни положительными, ни упорядоченными. Эти свойства обеспечиваются специальной последующей обработкой.

Сама итерационная процедура представляет собой  $QR$ -алгоритм Фрэнсиса, адаптированный Голубом и Райншем к задаче вычисления сингулярных чисел. Шаг алгоритма устроен следующим образом. Исходя из  $B_k$ , алгоритм определяет собственные значения  $\lambda_1$  и  $\lambda_2$  нижней угловой  $2 \times 2$ -подматрицы матрицы  $B_k^T B_k$ ; в качестве сдвига  $\sigma_k$  берется то  $\lambda_i$ , которое ближе к значению последнего диагонального элемента матрицы  $B_k^T B_k$ . Строится ортогональная матрица  $V_k$  так, чтобы произведение

$$V_k^T (B_k^T B_k - \sigma_k I_n) \quad (18.11)$$

было верхней треугольной матрицей. Строится ортогональная матрица  $U_k$  так, чтобы произведение

$$B_{k+1} = U_k^T B_k V_k \quad (18.12)$$

было верхней двухдиагональной матрицей.

Численная реализация процесса отличается от прямолинейного использования этих формул. В частности, матрица  $B_k^T B_k$  не формируется, и сдвиг  $\sigma_k$  выполняется неявным образом.

Чтобы упростить обозначения, опустим верхние индексы, указывающие номер итерации  $k$ . В обозначениях формулы (18.7) нижняя угловая  $2 \times 2$ -подматрица матрицы  $B^T B$  выглядит так:

$$\begin{bmatrix} q_{n-1}^2 + e_{n-1}^2 & e_n q_{n-1} \\ e_n q_{n-1} & q_n^2 + e_n^2 \end{bmatrix}.$$

Ее характеристическое уравнение имеет вид

$$(q_{n-1}^2 + e_{n-1}^2 - \lambda)(q_n^2 + e_n^2 - \lambda) - (e_n q_{n-1})^2 = 0. \quad (18.13)$$

Поскольку нам нужен корень уравнения (18.13), ближайший к  $q_n^2 + e_n^2$ , то удобно сделать подстановку

$$\delta = q_n^2 + e_n^2 - \lambda. \quad (18.14)$$

Для  $\delta$  получаем

$$\delta^2 + (q_{n-1}^2 - q_n^2 + e_{n-1}^2 - e_n^2)\delta - (e_n q_{n-1})^2 = 0. \quad (18.15)$$

Решение уравнения (18.15) упрощается, если положить

$$f = \frac{q_n^2 - q_{n-1}^2 + e_n^2 - e_{n-1}^2}{2e_n q_{n-1}}, \quad (18.16)$$

$$\gamma = \frac{\delta}{e_n q_{n-1}}. \quad (18.17)$$

Тогда  $\gamma$  удовлетворяет уравнению

$$\gamma^2 - 2f\gamma - 1 = 0. \quad (18.18)$$

Его корень  $\hat{\gamma}$  с меньшим модулем выражается формулой

$$\hat{\gamma} = 1/t, \quad (18.19)$$

где

$$t = \begin{cases} [-f - (1 + f^2)^{1/2}], & f \geq 0, \\ [-f + (1 + f^2)^{1/2}], & f < 0. \end{cases} \quad (18.20)$$

Теперь из (18.14) и (18.17) находим, что корнем уравнения (18.13), ближайшим к  $q_n^2 + e_n^2$ , является

$$\hat{\lambda} = q_n^2 + e_n^2 - \hat{\gamma} e_n q_{n-1} = q_n^2 + e_n(e_n - q_{n-1}/t), \quad (18.21)$$

а сдвиг определяется как

$$\sigma = \hat{\lambda}. \quad (18.22)$$

Далее нужно построить  $V$  таким образом, чтобы матрица  $V^T(B^TB - \sigma I)$  была верхней треугольной (см. (18.11)). Заметим, что трехдиагональная форма  $B^TB$  обеспечивает трехдиагональный вид матриц  $V^T(B^TB - \sigma I)V$  и  $V^T(B^TB)V$ .

Имеется частичное обращение этих утверждений, которое приводит к алгоритму, реально используемому для вычисления  $V$ .

**Теорема 18.23 (перефразированный вариант теоремы из [59]).** Пусть  $B^TB$  — трехдиагональная матрица с ненулевыми поддиагональными элементами,  $V$  — ортогональная матрица,  $\sigma$  — произвольный скаляр и, кроме того, матрица

$$V^T(B^TB)V \quad (18.24)$$

трехдиагональная; поддиагональные элементы первого столбца матрицы

$$V^T(B^TB - \sigma I) \quad (18.25)$$

равны нулю. В таком случае матрица  $V^T(B^TB - \sigma I)$  верхняя треугольная.

Исходя из этой теоремы, матрицы  $V$  и  $U$  в (18.11) и (18.12) будут вычисляться как произведения вращений Гивенса

$$V = R_1 \dots R_{n-1}, \quad (18.26)$$

$$U^T = T_{n-1} \dots T_1. \quad (18.27)$$

Здесь  $R_i$  оперирует со столбцами  $i$  и  $i+1$ , а  $T_i$  — со строками  $i$  и  $i+1$  матрицы  $B$ .

Первое вращение  $R_1$  определяется так, чтобы удовлетворить условие (18.25) теоремы 18.23. Остальные вращения  $R_i$  и  $T_i$  выбираются так, чтобы удовлетворить условие (18.24) и не нарушить условие (18.25).

Заметим, что как раз использованием матриц  $T_i$  этот алгоритм отличается от стандартного неявного QR-алгоритма для симметричных матриц (см. алгоритм Мартина и Уилкинсона в книге [8]). В случае симметричной задачи имеется трехдиагональная симметричная матрица, скажем  $Y$ , и нужно построить матрицу

$$\tilde{Y} = V^T Y V. \quad (18.28)$$

В описываемом же алгоритме  $Y$  находится в факторизованной форме  $B^TB$ , где  $B$  — двухдиагональная матрица. Матрицу  $\tilde{Y}$  также нужно построить в факторизованной форме:  $\tilde{Y} = \tilde{B}^T \tilde{B}$ . Следовательно,  $\tilde{B}$  должна быть матрицей вида  $\tilde{B} = U^T B V$ , где  $V$  — та же ортогональная матрица, что и в (18.28), и  $U$  — также ортогональная матрица.

Переходя теперь к вычислению вращений  $R_i$  и  $T_i$  из формул (18.26) и (18.27), заметим, что первый столбец матрицы  $B^TB - \sigma I$  имеет вид

$$\begin{bmatrix} q_1^2 - \sigma \\ q_1 e_2 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \quad (18.29)$$

где  $\sigma$  вычисляется в соответствии с (18.22). Первое вращение  $R_1$  определяется требованием, чтобы второй элемент первого столбца матрицы  $R_1^T(B^TB - \sigma I)$  был нулем.

Остальные вращения определяются в порядке  $T_1, R_2, T_2, \dots, R_{n-1}, T_{n-1}$  и применяются в порядке, указываемом скобками в следующем выражении:

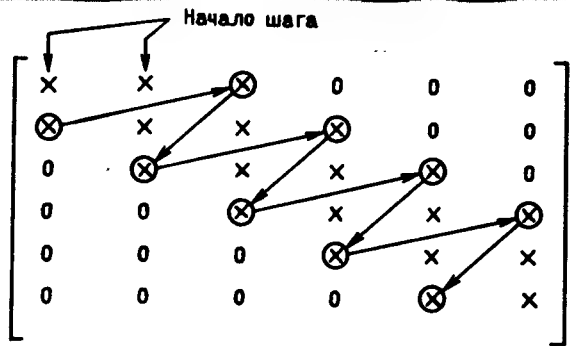
$$T_{n-1}(\dots T_2((T_1(BR_1))R_2)\dots R_{n-1}). \quad (18.30)$$

Вращение  $T_i$ ,  $i = 1, \dots, n-1$ , оперирует со строками  $i$  и  $i+1$  и аннулирует элемент в позиции  $(i+1, i)$ . Вращение  $R_i$ ,  $i = 1, \dots, n-1$ , оперирует со столбцами  $i$  и  $i+1$  и при  $i \geq 2$  аннулирует элемент в позиции  $(i-1, i+1)$ .

Это пример алгоритмического процесса, называемого иногда вытеснением. Рис. 18.1 показывает для  $n = 6$  последовательность появления и исчезновения элементов во внешности двухдиагональной полосы.

Из предыдущего обсуждения очевидно, что вычисление сингулярного разложения произвольной  $m \times n$ -матрицы можно свести к более специальной задаче сингулярного разложения невырожденной двухдиагональной





Р и с. 18.1. Иллюстрация процесса вытеснения в одном QR-шаге для случая  $n = 6$

матрицы  $B$  в форме (18.7). Сейчас будет сформулирован алгоритм QRBD выполняющий один шаг этой основной части в вычислении сингулярного разложения. Прежде всего алгоритм сравнивает внедиагональные элементы  $e_i$  с заданным допуском  $\epsilon$ . Если имеются индексы  $i$ ,  $2 \leq i \leq n$ , для которых  $|e_i| < \epsilon$ , то наибольший такой индекс помещается в ячейку  $l$ , и алгоритм прекращает работу. В противном случае выполняется присваивание  $l = 1$ ; такое значение  $l$  указывает, что  $|e_i| \geq \epsilon$  для  $2 \leq i \leq n$ . После этого алгоритм переходит к вычислению сдвига (см. (18.22)). Далее алгоритм вычисляет и применяет вращения  $R_i$  и  $T_i$ ,  $i = 1, \dots, n-1$  (см. (18.30)). Когда алгоритм заканчивает работу, элементы преобразованной матрицы  $\tilde{B} = U^T B V$  замещают в памяти одноименные элементы  $B$ .

В некоторых случаях, например при решении задачи НК, нужно умножать некоторые другие векторы или матрицы на матрицы  $U^T$  и  $V$ . Алгоритм QRBD предоставляет пользователю возможность задать  $k \times n$ -матрицу  $W = \{w_{ij}\}$  и  $n \times p$ -матрицу  $G = \{g_{ij}\}$ , которые будут заменены произведениями  $WV$  и  $U^T G$  соответственно.

Алгоритм 18.31. QRBD ( $q, e, n, \epsilon, l, W, k, G, p$ ):

1. Комментарий. Шаги 2 и 3 предназначены для выявления малых внедиагональных элементов.
2. Для  $i := n, n-1, \dots, 2$  выполнить шаг 3.
3. Если  $|e_i| \leq \epsilon$ , положить  $l := i$  и перейти к шагу 14.
4. Положить  $l := 1$  и вычислить  $\sigma$  в соответствии с формулами (18.16) – (18.22).
5. Положить  $i := 2$ .
6. Положить  $e_1 := q_1 - \sigma/q_1$  и  $z := e_2$  (см. (18.29)).
7. Выполнить алгоритм G1( $e_{l-1}, z, c, s, e_{l-1}$ ).
8. Выполнить алгоритм G2( $c, s, q_{l-1}, e_l$ ) и для  $j := 1, \dots, k$  выполнить алгоритм G2( $c, s, w_{j, l-1}, w_{jl}$ ).
9. Положить  $z := sq_l$ ,  $q_l := cq_l$ .
10. Выполнить алгоритм G1( $q_{l-1}, z, c, s, q_{l-1}$ ).
11. Выполнить алгоритм G2( $c, s, e_l, q_l$ ) и для  $j := 1, \dots, p$  выполнить алгоритм G2( $c, s, g_{l-1, j}, g_{lj}$ ).

12. Если  $i = n$ , перейти к шагу 14.

13. Положить  $z := se_{l+1}$ ,  $e_{l+1} := ce_{l+1}$ ,  $i := i + 1$  и перейти к шагу 7.

14. Комментарий. Если  $l = 1$ , то был выполнен один полный QR-шаг для двухдиагональной матрицы. Если  $l > 1$ , то элемент  $e_l$  мал, и матрицу можно расщепить в этом месте.

Повторным применением алгоритма QRBD к невырожденной верхней двухдиагональной матрице  $B$  строится последовательность двухдиагональных матриц  $B_k$  с диагональными элементами  $q_1^{(k)}, \dots, q_n^{(k)}$  и наддиагональными элементами  $e_2^{(k)}, \dots, e_n^{(k)}$ . Согласно теореме 18.5, произведение  $e_n^{(k)} q_n^{(k)}$  квадратично сходится к нулю при  $k \rightarrow \infty$ . Из предположения о невырожденности  $B$  вытекает, что последовательность  $\{q_n^{(k)}\}$  отделена от нуля. Следовательно, последовательность  $\{e_n^{(k)}\}$  квадратично сходится к нулю. На практике итерации прекращают, когда  $|e_n^{(k)}| < \epsilon$ .

После того как величина  $|e_n^{(k)}|$  признана достаточно малой, переходят к вычислению сингулярного разложения для двухдиагональной матрицы (или нескольких таких матриц) порядка  $n-1$  (или меньшего); это делается так же, как и выше, с заменой  $n$  на  $n-1$ . Так происходит, пока не будет  $n = 1$ . Не более чем  $n-1$  повторений описанной процедуры приводят к разложению  $B = \tilde{U} \tilde{S} \tilde{V}^T$ .

Практика показывает, что при значениях параметра точности  $\eta$  в пределах примерно от  $10^{-18}$  до  $10^{-8}$  и при  $\epsilon = \eta \|B\|$  обычно бывает достаточно около  $2n$  повторений алгоритма QRBD, чтобы все внедиагональные элементы стали по модулю меньше  $\epsilon$ .

Диагональные элементы  $\tilde{S}$  в общем случае не являются неотрицательными или невозрастающими. Чтобы поправить этот недостаток, возьмем диагональную матрицу  $D$ , на диагонали которой стоят плюс и минус единицы, выбранные так, что у матрицы

$$\hat{S} = \tilde{S} D \quad (18.32)$$

диагональные элементы неотрицательны. Далее выбираем матрицу перестановки  $P$  так, чтобы диагональные элементы матрицы

$$S = P^T \hat{S} P \quad (18.33)$$

не возрастали. Легко видеть, что равенство

$$B = (\tilde{U} P)(P^T \hat{S} P)(\tilde{V} D P)^T \equiv \hat{U} S \hat{V}^T$$

представляет собой сингулярное разложение матрицы  $B$ , причем диагональные элементы  $S$  неотрицательны и не возрастают.

Матрица  $V$  из (18.9) будет вычислена как произведение

$$V = H_2 \dots H_n R_1 \dots R_p D P. \quad (18.34)$$

Здесь  $H_i$  — преобразования Хаусхолдера из формулы (18.6);  $R_i$  — правые вращения Гивенса, которые строились и использовались перед QR-шагами и в самих этих шагах (см. (18.30) и текст, следующий за (18.10)); матрицы  $D$  и  $P$  определены равенствами (18.32) и (18.33) соответственно.

Аналогично матрица  $U$  из (18.9) строится как произведение

$$U^T = PT_\mu \dots T_1 Q_n \dots Q_1, \quad (18.35)$$

где  $P$  определена в (18.33);  $T_i$  — левые вращения Гивенса, которые строятся перед  $QR$ -шагами и в самих этих шагах (см. (18.10) и (18.30));  $Q_i$  — преобразования Хаусхолдера из формулы (18.6).

#### § 4. Решение задачи НК посредством сингулярного разложения

Сингулярное разложение матрицы  $A$

$$A_{m \times n} = U_{m \times m} \begin{bmatrix} S_{n \times n} \\ 0_{(m-n) \times n} \end{bmatrix} V_{n \times n}^T \quad (18.36)$$

можно использовать для сведения задачи  $Ax \cong b$  к эквивалентной задаче

$$\begin{bmatrix} S \\ 0 \end{bmatrix} p \cong q, \quad (18.37)$$

где

$$g \equiv \begin{bmatrix} g^{(1)} \\ g^{(2)} \end{bmatrix} \begin{matrix} \} n \\ \} m-n \end{matrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} b = U^T b, \quad (18.38)$$

$$x = Vp. \quad (18.39)$$

Может оказаться полезным рассмотрение последовательности пробных решений  $x^{(k)}$ , определяемых формулами

$$x^{(k)} = \sum_{j=1}^k p_j v_j, \quad (18.40)$$

где  $v_j$  —  $j$ -й столбец  $V$ .

Заметим, что векторы  $x^{(k)}$  удобно вычислять так:

$$x^{(0)} = 0, \quad (18.41)$$

$$x^{(k)} = x^{(k-1)} + p_k v_k, \quad k = 1, \dots, n. \quad (18.42)$$

Норма  $\rho_k$  невязки, отвечающей пробному решению  $x^{(k)}$ , определяется как  $\rho_k = \|Ax^{(k)} - b\|$  и удовлетворяет соотношению

$$\rho_k^2 = \|g^{(2)}\|^2 + \sum_{i=k+1}^n (g_i^{(1)})^2. \quad (18.43)$$

Числа  $\rho_k$  можно вычислять следующим образом:

$$\rho_n^2 = \|g^{(2)}\|^2, \quad (18.44)$$

$$\rho_k^2 = \rho_{k+1}^2 + (g_{k+1}^{(1)})^2, \quad k = n-1, \dots, 0. \quad (18.45)$$

Здесь  $g^{(1)}, g^{(2)}$  — подвекторы вектора  $g$ , определенные в (18.38).

Столбцы  $V$ , соответствующие малым сингулярным числам, можно интерпретировать как показатели приближенной линейной зависимости между столбцами  $A$ . Это было отмечено в гл. 12. Дальнейшие замечания по поводу практического использования сингулярных чисел и величин  $\rho_k$  приведены в § 6 гл. 25.

#### § 5. Организация программы, вычисляющей сингулярное разложение

Опишем способ выполнения сингулярного разложения с экономным расходом памяти. Будем считать, что  $m \geq n$ . Замечания о случае  $m < n$  см. в § 1.

Прежде всего, если  $mn$  велико и  $m \geq n$ , то мы можем вначале привести матрицу исходных данных  $[A : b]$  к верхней треугольной матрице порядка  $n+1$ ; это делается посредством последовательности преобразований Хаусхолдера (см. гл. 27). Затем в соответствии с (18.6) матрица  $A$  приводится к двухдиагональному виду  $\begin{bmatrix} B \\ 0 \end{bmatrix}$ . Ненулевые элементы  $B$  замещают соответствующие элементы  $A$  в массиве с именем  $A$ .

Преобразования  $Q_i$  из (18.6) можно применять к вектору  $b$  по мере их формирования; хранить их ненужно. Вектор, полученный в результате преобразований, замещает  $b$  в массиве с именем  $b$ . Преобразования  $H_i$  из (18.6) размещают в освобождающихся позициях верхнего треугольника  $A$  и дополнительном массиве длины  $n$ ; назовем этот массив  $h$ .

Когда приведение к двухдиагональному виду закончено, ненулевые элементы  $B$  переписываются в два массива длины  $n$ :  $q$  (позиции  $q_1, \dots, q_n$ ) и  $e$  (позиции  $e_2, \dots, e_n$ ); ячейка  $e_1$  используется  $QR$ -алгоритмом как рабочая.

Вычисление матрицы  $V$  из формулы (18.36) начинается с формирования в явном виде произведения  $H_2 \dots H_n$ , входящего в (18.34). Это вычисление можно организовать таким образом, чтобы полученное произведение занимало первые  $n$  строк массива  $A$ ; дополнительных массивов не требуется.

$QR$ -алгоритм применяется к матрице  $B$ , хранимой парой массивов  $q$  и  $e$ . Каждое очередное вращение  $R_i$ , построенное в  $QR$ -алгоритме, умножается на частичное произведение, хранимое в массиве  $A$  в качестве будущей матрицы  $V$  (см. (18.34)). Аналогично каждое вращение  $T_i$  умножается на вектор, хранимый в массиве  $b$  в качестве будущего вектора  $U^T b$  (см. (18.35) и (18.38)).

По окончании  $QR$ -итераций в ячейках  $e_i$ ,  $i = 2, \dots, n$ , будут храниться малые числа. Теперь нужно сделать неотрицательными, а затем упорядочить числа, находящиеся в ячейках  $q_i$ ,  $i = 1, \dots, n$ . Применение к массиву  $A$  соответствующих перемен знаков и перестановок завершает вычисление матрицы  $V$  (см. (18.34)). Применение перестановок к массиву  $b$  завершает вычисление вектора  $g = U^T b$  (см. (18.35) и (18.38)).

При необходимости можно вычислить в соответствии с (18.41) и (18.42) пробные решения и хранить их в качестве столбцов верхней  $n \times n$ -подматрицы массива  $A$ , где прежде хранилась матрица  $V$ .

Если сингулярное разложение вычисляется для того, чтобы достигнуть лучшего понимания конкретной задачи наименьших квадратов, то полезно иметь программу, которая бы печатала в удобном формате различные величины, извлекаемые из сингулярного разложения.

# Упражнения

18.46. Доказать, что если  $A_1$  — симметричная трехдиагональная матрица, то таковы же все матрицы  $A_k$ ,  $k = 2, \dots$ , формулы (18.3).

18.47. Доказать, что специальный QR-алгоритм, определяемый формулами (18.11), (18.12) и (18.22), сходится в одну итерацию в случае двухдиагональной матрицы порядка 2.

## ГЛАВА 19

### ДРУГИЕ МЕТОДЫ ДЛЯ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ

В качестве вычислительного аппарата для решения задач теории наименьших квадратов мы выделили ортогональные преобразования Хаусхолдера. При таком подходе численная устойчивость, характерная для ортогональных преобразований (см. гл. 15–17), сочетается с гибкостью, позволяющей легко приспосабливаться к специальным ситуациям, например последовательному накоплению данных (гл. 27). Мы предложили также использовать сингулярное разложение как средство достигнуть лучшего понимания плохо обусловленных задач (гл. 18, 25).

В этой главе будут описаны некоторые другие численные методы для задачи наименьших квадратов. В табл. 19.1 приведены старшие члены формул для числа операций при решении задачи  $A_m \times n x \cong b$ ,  $m > n$ , различными методами.

Мы обсудим, в частности, методы, основанные на математической эквивалентности задачи наименьших квадратов  $Ax \cong b$  и системы линейных уравнений  $(A^T A)x = (A^T b)$ . Эту систему называют *системой нормальных уравнений* исходной задачи. Методы, предполагающие формирование и решение нормальных уравнений, требуют обычно примерно вдвое меньше операций, чем алгоритм Хаусхолдера. Однако, чтобы получить этими методами решения того же качества, что в алгоритме Хаусхолдера при относительной точности  $\eta$ , здесь требуется работать с точностью  $\eta^2$ .

В большинстве специальных ситуаций, например в случае последовательного накопления данных (см. гл. 27), можно организовать оба метода — алгоритм нормальных уравнений и алгоритм Хаусхолдера — так, что требуется примерно одно и то же число единиц хранения. Заметим, однако, что если в качестве единицы хранения принимается машинное слово одинаковой для обоих методов длины, то алгоритм Хаусхолдера сможет обрабатывать более широкий класс задач, чем алгоритм нормальных уравнений.

Второй метод, обсуждаемый в данной главе, — это *модифицированная ортогонализация Грама–Шмидта* (MGS)\*. Численные свойства этого метода очень сходны со свойствами алгоритма Хаусхолдера. Факторизованное представление матрицы  $Q$ , используемое в алгоритме Хаусхолдера, требует приблизительно на  $n^2/2$  меньше ячеек памяти, чем явное представление, используемое в MGS. Как показывает наш опыт, именно по этой причине алгоритм Хаусхолдера легче, чем MGS, адаптируется к различным специальным приложениям.

\* MGS — сокращение английского словосочетания *Modified Gram – Schmidt Orthogonalization*. (Примеч. пер.)

Таблица 19.1

Количество операций для различных численных методов решения задачи наименьших квадратов

| Метод                                                                         | Асимптотическое число операций *) |
|-------------------------------------------------------------------------------|-----------------------------------|
| Приведение к треугольному виду методом Хаусхолдера                            | $mn^2 - n^3/3$                    |
| Сингулярное разложение:                                                       |                                   |
| прямое применение к $A$                                                       | $2mn^2 + \sigma(n) **$            |
| приведение $A$ к треугольной матрице $R$                                      |                                   |
| методом Хаусхолдера в сингулярное разложение $R$                              | $mn^2 + 5n^3/3 + \sigma(n) **$    |
| Формирование нормальных уравнений                                             | $mn^2/2$                          |
| Решение нормальных уравнений методом Холесского                               | $n^3/6$                           |
| Решение нормальных уравнений методом Гаусса–Жордана (для пошаговой регрессии) | $n^3/3$                           |
| Спектральный анализ нормальных уравнений                                      | $4n^3/3 + \sigma(n) **$           |
| Метод Грама–Шмидта (классический или модифицированный)                        | $mn^2$                            |

\*) Под операцией понимается пара "умножить (или разделить) и сложить".

\*\*) Член  $\sigma(n)$  учитывает итерационную фазу в вычислении сингулярных чисел или собственных значений. Если считать, что для сходимости QR-алгоритма нужно примерно  $2n$  шагов, то  $\sigma(n)$  равно приблизительно  $4n^2$ .

В качестве конкретного примера этой особенности компактного хранения заметим, что с помощью преобразований Хаусхолдера можно вычислить решение  $\hat{x}$  задачи  $Ax \cong b$  и вектор невязки  $r = b - A\hat{x}$ , используя  $m(n+1) + 2n + m$  ячеек памяти. Как в алгоритме нормальных уравнений, так и в алгоритме MGS требуются дополнительные  $n(n+1)/2$  ячеек. При этом, как отмечалось выше, оценка запросов к памяти должна принимать во внимание, что метод нормальных уравнений требует  $\eta^2$ -разрядности для обработки того же класса задач, который обрабатывается алгоритмами Хаусхолдера и MGS в арифметике с относительной точностью  $\eta$ .

### § 1. Нормальные уравнения и разложение Холесского

В книгах, ограничивающихся кратким разбором задачи наименьших квадратов, чаще всего рекомендуют именно метод нормальных уравнений. Обе части исходной задачи  $A_m \times n x \cong b$  умножают слева на матрицу  $A^T$ . В результате получается система

$$P_{n \times n} x = d_{n \times 1}, \quad (19.1)$$

называемая *системой нормальных уравнений для данной задачи*. Здесь

$$P = A^T A, \quad (19.2)$$

$$d = A^T b. \quad (19.3)$$

Уравнение (19.1) можно было бы вывести непосредственно из условия, что для искомого решения вектор невязки  $b - Ax$  должен быть ортогонален к пространству столбцов матрицы  $A$ . Это условие записывается уравнением

$$A^T(b - Ax) = 0, \quad (19.4)$$

которое, как видно из (19.2) и (19.3), эквивалентно уравнению (19.1).

Хотя  $P$  имеет тот же ранг, что и  $A$ , и, следовательно, может быть вырождена, уравнение (19.1) всегда совместно. Чтобы убедиться в этом, заметим, что, согласно (19.3), вектор  $d$  принадлежит пространству строк  $A$ . Но, как показывает (19.2), пространство строк  $A$  совпадает с пространством столбцов  $P$ .

Систему (19.1) можно решать любым методом, предназначенным для квадратных совместных систем линейных уравнений. Из (19.2) следует, однако, что матрица  $P$  симметрична и неотрицательно определена, что позволяет использовать разложение Холесского с его превосходными качествами экономичности и устойчивости [7].

Метод Холесского, который мы сейчас опишем, основан на факте существования верхней треугольной (действительной)  $n \times n$ -матрицы  $U$  такой, что

$$U^T U = P. \quad (19.5)$$

Поэтому решение  $x$  уравнения (19.1) можно получить, решая две треугольные системы:

$$U^T y = d, \quad (19.6)$$

$$Ux = y. \quad (19.7)$$

Иначе процесс решения (19.6) относительно  $y$  можно оформить как часть разложения Холесского подходящей окаймленной матрицы. С этой целью положим  $\tilde{A} = [A : b]$  и

$$\tilde{P} = \tilde{A}^T \tilde{A} \equiv \begin{bmatrix} P & d \\ d^T & \omega^2 \end{bmatrix}. \quad (19.8)$$

Заметим, что  $P$  и  $d$  в (19.8) обозначают те же матрицу и вектор, что и в формулах (19.2) и (19.3), а для числа  $\omega^2$  справедливо  $\omega^2 = b^T b$ . Разложение Холесского матрицы  $\tilde{P}$  имеет вид

$$\tilde{P} = \tilde{U}^T \tilde{U}, \quad (19.9)$$

где

$$\tilde{U} = \begin{bmatrix} U & y \\ \underline{0} & \underline{\rho} \end{bmatrix} \begin{matrix} \} n \\ \} 1 \end{matrix}, \quad (19.10)$$

причем  $U$  и  $y$  в (19.10) удовлетворяют соотношениям (19.5) и (19.6). Кроме того, как легко проверить, для числа  $\rho$  в (19.10)  $|\rho| = \|b - A\hat{x}\|$ , где  $\hat{x}$  — любой вектор, минимизирующий  $\|b - A\hat{x}\|$ .

Для удобства читателя мы опишем, исходя из (19.5), детали вычисления разложения Холесского. Очевидно, что алгоритм непосредственно приложим и к (19.9).

Из (19.5) получаем равенства

$$\sum_{k=1}^i u_{ki} u_{kj} = p_{ij}, \quad (19.11)$$

$$i = 1, \dots, n;$$

$$j = i, \dots, n.$$

Разрешая относительно  $u_{ij}$  уравнение с правой частью  $p_{ij}$ , приходим к следующим соотношениям, составляющим алгоритм факторизации Холесского (называемый также методом квадратных корней или методом Банахевича):

$$v_i = p_{ii} - \sum_{k=1}^{i-1} u_{ki}^2, \quad (19.12)$$

$$u_{ii} = v_i^{1/2}, \quad (19.13)$$

$$u_{ij} = \frac{p_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj}}{u_{ii}}, \quad (19.14)$$

$$j = i+1, \dots, n, \quad i = 1, \dots, n.$$

В формулах (19.12) и (19.14) суммы считаются равными нулю при  $i = 1$ .

Теоретически все  $v_i > 0$ ,  $i = 1, \dots, n$ , если  $\text{rang } A = n$ . Уравнения (19.12)–(19.14) в этом случае определяют значение каждого  $u_{ij}$  однозначно. Если, однако,  $\text{rang } A = k < n$ , то найдется первое значение  $i$ , скажем  $i = t$ , для которого  $v_t = 0$ . Тогда при  $i = t$  числители в правых частях формул (19.14) также должны быть нулями для  $j = t+1, \dots, n$  — ведь система уравнений (19.11) совместна. В таком случае имеется некоторая свобода в выборе значений для элементов  $u_{ij}$ ,  $j = t+1, \dots, n$ . Всегда допустимым является набор значений  $u_{ij} = 0$ ,  $j = t+1, \dots, n$ , (см. упражнение 19.37).

Из сказанного следует, что теоретически алгоритм (19.12)–(19.14) дает решение уравнения (19.5) при любом ранге  $A$ , если внести в него следующую модификацию: (19.14) заменяется на

$$u_{ij} = 0, \quad j = i+1, \dots, n, \quad (19.15)$$

для любого значения  $i$ , при котором  $u_{ii} = 0$ .

Заметим, что верхняя треугольная матрица  $R$  в разложении Хаусхолдера  $QA = \begin{bmatrix} R \\ 0 \end{bmatrix}$  удовлетворяет соотношению  $R^T R = P$ . Если  $\text{rang } A = n$ , то решение (19.5) единственно с точностью до знаков строк  $U$  (см. упражнение 2.17). Поэтому в случае  $\text{rang } A = n$  матрица  $R$  разложения Хаусхолдера совпадает с точностью до знаков строк с матрицей Холесского  $U$ .

Разложение Холесского можно вычислять и в виде

$$L^T L = P, \quad (19.16)$$

где  $L$  — нижняя треугольная  $n \times n$ -матрица. Формулы для элементов  $L$

таковы:

$$l_{ii} = \left( p_{ii} - \sum_{k=i+1}^n l_{ki}^2 \right)^{1/2}, \quad (19.17)$$

$$l_{ij} = \frac{p_{ij} - \sum_{k=i+1}^n l_{ki} l_{kj}}{l_{ii}}, \quad (19.18)$$

$$j = 1, \dots, i-1, \quad i = n, \dots, 1.$$

При реальных вычислениях появляется возможность, что в разложении Холесского для (теоретически) неотрицательно определенной матрицы  $P$  значение  $u_i$ , вычисленное по формуле (19.12), будет отрицательным вследствие округлений. Это может быть результатом накопления погрешностей в вычислениях по формулам (19.2) и (19.12)–(19.14).

Если для некоторого  $i$  вычисленное значение  $u_i$  отрицательно, то одна из возможностей продолжить вычисление состоит в том, чтобы положить  $u_i = 0$ , а затем прибегнуть к (19.15). В сущности, та же идея заложена в фортранную подпрограмму из [91].

Другой возможный способ – выполнить симметричную перестановку строк и столбцов, максимизирующую значение  $u_i$  на  $i$ -м шаге алгоритма; информацию о перестановке нужно сохранить. Эффект будет тот, что неположительные значения  $u_i$  могут появиться лишь после обработки всех положительных значений. Если все остающиеся значения  $u_i$  неположительны, то соответствующие строки  $U$  можно взять нулевыми.

Реализация такой стратегии перестановок требует некоторого переупорядочения операций в алгоритме (19.12)–(19.14), с тем чтобы при выборе главного элемента имелись частично вычисленные значения  $u_i$ . Если перестановки применяются к окаймленной матрице (19.8), то последние строка и столбец не должны принимать в них участия.

Разложение Холесского для матрицы (19.8) приводит к верхней треугольной матрице (19.10), имеющей ту же связь с задачей НК, что и треугольная матрица, полученная из  $\tilde{A}$  посредством триангуляризации Хаусхолдера. Это замечание может оказаться полезным в случае, когда исходные данные уже имеют форму нормальных уравнений и нужно провести анализ задачи с помощью сингулярного разложения.

Если игнорировать эффекты округлений, то информация, предоставляемая сингулярным разложением  $A$  (см. гл. 18 и § 5 гл. 25), может быть получена и из спектрального анализа матрицы  $P$  (см. (19.2)). Именно, если имеем спектральное разложение  $P = VS^2V^T$  с упорядоченными собственными значениями  $s_{11}^2 \geq \dots \geq s_{nn}^2$ , то можно вычислить  $g^{(1)} = S^{-1}V^Td$ ,  $\gamma = (\omega^2 - \|g^{(1)}\|^2)^{1/2}$ , где  $d$  и  $\omega^2$  определены равенством (19.8). При замене переменных  $x = Vp$  исходная задача НК переходит в задачу

$$\begin{bmatrix} S \\ 0 \end{bmatrix} p \cong \begin{bmatrix} g^{(1)} \\ \gamma \end{bmatrix}. \quad (19.19)$$

Эта задача наименьших квадратов эквивалентна задаче (18.37), полученной посредством сингулярного разложения: диагональная матрица  $S$  и  $n$ -вектор

$g^{(1)}$  в (19.19) те же, что и в (18.37);  $\gamma$  из (19.19) и  $g^{(2)}$  из (18.37) удовлетворяют соотношению  $\gamma = \|g^{(2)}\|$ .

Однако при фиксированной разрядности машинного слова величины  $S$ ,  $g^{(1)}$  и  $\gamma = \|g^{(2)}\|$  будут определены путем спектрального анализа матрицы  $[A : b]^T [A : b]$  менее точно, чем если бы применялся сингулярный анализ к самой матрице  $\tilde{A} = [A : b]$  или треугольной матрице  $R$ , полученной из  $\tilde{A}$  алгоритмом Хаусхолдера.

Как отмечено во введении к данной главе, при фиксированной относительной точности  $\eta$  машинной арифметики преобразования Хаусхолдера позволяют решать более широкий класс задач, чем если бы формировались нормальные уравнения, а затем применялось разложение Холесского. Это утверждение можно проиллюстрировать следующим примером. Рассмотрим  $3 \times 2$ -матрицу

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 - \epsilon \end{bmatrix}.$$

Предположим, что значение  $\epsilon$  существенно для данной задачи, например  $\epsilon > 100\eta$ , но  $\epsilon^2 < \eta$ , так что при вычислениях с относительной точностью  $\eta$  будет  $1 - \epsilon \neq 1$ , но  $3 + \epsilon^2$  вычисляется как 3. Поэтому вместо

$$A^T A = \begin{bmatrix} 3 & 3 - \epsilon \\ 3 - \epsilon & 3 - 2\epsilon + \epsilon^2 \end{bmatrix}$$

будет вычислена (с точностью до случайных погрешностей, не превосходящих по абсолютной величине  $3\eta$ ) матрица

$$\begin{bmatrix} 3 & 3 - \epsilon \\ 3 - \epsilon & 3 - 2\epsilon \end{bmatrix}.$$

Последующее вычисление верхней треугольной матрицы

$$\begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix},$$

для которой  $R^T R = A^T A$ , проводится в соответствии с методом Холесского (19.12)–(19.14) по формулам

$$r_{11} = \sqrt{3}, \quad r_{12} = \frac{3 - \epsilon}{\sqrt{3}}, \quad r_{22}^2 = p_{22} - r_{12}^2.$$

При вычислении  $r_{22}$  (опять же с точностью до случайных погрешностей произвольного знака, не превышающих по модулю  $3\eta$ ) получим

$$r_{22}^2 = 3 - 2\epsilon - \frac{9 - 6\epsilon}{3} = 0.$$

Соответствующий точный результат был бы, конечно, равен

$$r_{22}^2 = 3 - 2\epsilon + \epsilon^2 - \frac{9 - 6\epsilon + \epsilon^2}{3} = \frac{2\epsilon^2}{3}.$$

Таким образом, при использовании арифметики с относительной точностью  $\eta$  в вычисленном элементе  $r_{22}$  нет значащих цифр. Следовательно, матрица  $R$  неотличима от вырожденной. Причиной не являются недостатки алгоритма Холесского. Скорее здесь нашел отражение тот факт, что столбцы матрицы  $A^T A$  так близки к параллельности (линейной зависимости), что отсутствие точной параллельности нельзя установить в арифметике с точностью  $\eta$ .

Сопоставим эту ситуацию с той, что имеет место при прямой (без предварительного формирования  $A^T A$ ) триангуляризации преобразованиями Хаусхолдера. По формулам (10.5)–(10.11) вычисляем

$$s = -\sqrt{3}, \quad u = \begin{bmatrix} 1 + \sqrt{3} \\ 1 \\ 1 \end{bmatrix}, \quad b = -\sqrt{3}(1 + \sqrt{3}) = -(3 + \sqrt{3}).$$

Теперь, умножая на  $I + b^{-1}uu^T$  второй столбец  $a_2$  матрицы  $A$ , получаем

$$t = b^{-1}(u^T a_2) = -\frac{3 + \sqrt{3} - \epsilon}{3 + \sqrt{3}} = \frac{\epsilon(3 - \sqrt{3}) - 6}{6},$$

$$\tilde{a}_2 = a_2 + tu = \begin{bmatrix} 1 \\ 1 \\ 1 - \epsilon \end{bmatrix} + \frac{\epsilon(3 - \sqrt{3}) - 6}{6} \begin{bmatrix} 1 + \sqrt{3} \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{\epsilon - 3}{\sqrt{3}} \\ \frac{\epsilon(3 - \sqrt{3})}{6} \\ -\frac{\epsilon(3 + \sqrt{3})}{6} \end{bmatrix}.$$

Важно отметить, что вторая и третья компоненты  $\tilde{a}_2$ , имеющие порядок  $\epsilon$ , вычислены как разности величин порядка единицы. Поэтому в арифметике с  $\eta$ -точностью эти компоненты не будут утрачены вследствие округлений.

Последний шаг триангуляризации Хаусхолдера состоит в замене второй компоненты  $\tilde{a}_2$  значением (со знаком минус) евклидовой нормы второй и третьей компонент. Этот шаг не вызывает вычислительных затруднений:

$$r_{22} = -\left\{ \left[ \frac{\epsilon(3 - \sqrt{3})}{6} \right]^2 + \left[ \frac{-\epsilon(3 + \sqrt{3})}{6} \right]^2 \right\}^{1/2} = \frac{-\epsilon\sqrt{6}}{3}.$$

Объединяя результаты, мы получаем (с точностью до абсолютных погрешностей, не превосходящих по модулю  $3\eta$ ) матрицу

$$R = \begin{bmatrix} -\sqrt{3} & (\epsilon - 3)/\sqrt{3} \\ 0 & -\epsilon\sqrt{6}/3 \end{bmatrix}.$$

Окончательный вывод из этого примера таков. Используя арифметику с  $\eta$ -точностью и применяя преобразования Хаусхолдера непосредственно к  $A$ , мы вычислили треугольную матрицу, которая очевидным образом не вырождена. В то же время формирование нормальных уравнений и последующее разложение Холесского привели к вырожденной матрице.

## § 2. Модифицированная ортогонализация Грама–Шмидта

Ортогонализация Грама–Шмидта — это классический математический метод решения следующей задачи. Дана линейно независимая система векторов  $\{a_1, \dots, a_n\}$ ; нужно построить ортогональную систему  $\{q_1, \dots, q_n\}$  с тем свойством, что для  $k = 1, \dots, n$  подсистема  $\{q_1, \dots, q_k\}$  порождает то же  $k$ -мерное подпространство, что и  $\{a_1, \dots, a_k\}$ . Классические формулы, выражающие  $q_j$  через  $a_j$  и ранее вычисленные векторы  $q_1, \dots, q_{j-1}$ , таковы:

$$q_1 = a_1, \quad (19.20)$$

$$q_j = a_j - \sum_{i=1}^{j-1} r_{ij} q_i, \quad j = 2, \dots, n, \quad (19.21)$$

где

$$r_{ij} = \frac{a_j^T q_i}{q_i^T q_i}. \quad (19.22)$$

Чтобы записать эти формулы в матричных обозначениях, определим  $A$  как матрицу со столбцами  $a_j$ ,  $Q$  как матрицу со столбцами  $q_j$  и  $R$  как верхнюю треугольную матрицу с единичными диагональными элементами и наддиагональными элементами, задаваемыми формулами (19.22). Тогда (19.20) и (19.21) можно переписать в виде

$$A = QR. \quad (19.23)$$

Отсюда ясно, что ортогонализацию Грама–Шмидта можно рассматривать как еще один метод разложения матрицы в произведение матрицы с ортогональными столбцами и треугольной матрицы.

Эксперименты свидетельствуют [157], что процесс (19.20)–(19.22) обладает значительно меньшей численной устойчивостью, чем некоторый его вариант, математически ему эквивалентный. Устойчивость этого варианта, называемого модифицированным методом Грама–Шмидта, была установлена Бюрком [23]; ему принадлежат приводимые ниже оценки (19.35) и (19.36).

Заметим, что значение скалярного произведения  $a_j^T q_i$  не изменится, если  $a_j$  заменить вектором вида

$$a_j - \sum_{k=1}^{i-1} \alpha_k q_k, \quad (19.24)$$

так как  $q_k^T q_i = 0$  для  $k \neq i$ . Исходя из этого, рекомендуется заменить (19.22) на

$$r_{ij} = \frac{a_j^{(i)T} q_i}{q_i^T q_i}, \quad (19.25)$$

где

$$a_j^{(i)} = a_j - \sum_{k=1}^{i-1} r_{kj} q_k. \quad (19.26)$$

Если поставить задачу — минимизировать за счет выбора чисел  $\alpha_k$  норму вектора (19.24), то, как легко проверить, вектор с минимальной длиной задается формулой (19.26). Таким образом, вектор  $a_j$  в (19.22) и вектор  $a_j^{(i)}$  в (19.25) связаны неравенством  $\|a_j^{(i)}\| \leq \|a_j\|$ . Большая численная устойчивость модифицированного алгоритма проистекает как раз из этого факта.

Количество вычислений и необходимая память не увеличиваются модификацией, так как вместо непосредственного использования формулы (19.26) векторы  $a_j^{(i)}$  можно вычислять рекурсивно. Кроме того, вектор  $a_j^{(i+1)}$  можно записывать на место, которое прежде занимал вектор  $a_j^{(i)}$ .

Этот алгоритм, за которым закрепилось название *модифицированный метод Грама-Шмидта* [157], описывается следующими уравнениями:

$$a_j^{(1)} = a_j, \quad j = 1, \dots, n, \quad (19.27)$$

$$q_i = a_i^{(i)}, \quad (19.28)$$

$$d_i^2 = q_i^T q_i; \quad (19.29)$$

$$r_{ij} = \frac{a_j^{(i)T} q_i}{d_i^2}, \quad (19.30)$$

$$a_j^{(i+1)} = a_j^{(i)} - r_{ij} q_i, \quad j = i+1, \dots, n, \quad i = 1, \dots, n. \quad (19.31)$$

Чтобы использовать MGS \*) для решения задачи НК, можно составить расширенную матрицу

$$\tilde{A} = [A : b]. \quad (19.32)$$

После этого MGS применяется к  $m \times (n+1)$ -матрице  $\tilde{A}$ ; в результате будет получено разложение

$$\tilde{A} = \tilde{Q} \tilde{R}, \quad (19.33)$$

где  $\tilde{R}$  — верхняя треугольная матрица с единичными диагональными элементами. Наддиагональные элементы  $\tilde{R}$  выражаются формулами (19.30). Векторы  $q_i$  из (19.28) являются столбцами  $m \times (n+1)$ -матрицы  $\tilde{Q}$ .

\*) См. примечание на стр. 92. (Примеч. пер.)

Можно еще составить диагональную матрицу  $\tilde{D}$  порядка  $n+1$  с диагональными элементами  $d_i$ ,  $i = 1, \dots, n+1$ , задаваемыми формулами (19.29).

Для последующего обсуждения, которое приведет к решению задачи НК, возьмем какую-либо ортогональную  $m \times m$ -матрицу  $Q$ , удовлетворяющую соотношению

$$Q_{m \times m} \begin{bmatrix} \tilde{D} \\ 0 \end{bmatrix} \begin{matrix} \} n+1 \\ \} m-n-1 \end{matrix} = \tilde{Q}_{m \times (n+1)}.$$

Тогда

$$\tilde{A} = Q \begin{bmatrix} \tilde{D} \\ 0 \end{bmatrix} \tilde{R} = Q \begin{bmatrix} DR & Dc \\ 0 & d_{n+1} \\ 0 & 0 \end{bmatrix}. \quad (19.34)$$

В последнем переходе были использованы клеточные представления матриц  $\tilde{D}$  и  $\tilde{R}$

$$\tilde{R} = \begin{bmatrix} R & c \\ 0 & 1 \end{bmatrix} \begin{matrix} \} n \\ \} 1 \end{matrix}, \quad \tilde{D} = \begin{bmatrix} D & 0 \\ 0 & d_{n+1} \end{bmatrix} \begin{matrix} \} n \\ \} 1 \end{matrix}.$$

Теперь для произвольного вектора  $x$  имеем  $\|Ax - b\|^2 = \|Q^T(Ax - b)\|^2 = \|D(Rx - c)\|^2 + d_{n+1}^2$ . Следовательно, минимальным значением величины  $\|Ax - b\|$  будет  $|d_{n+1}|$ , и это значение достигается на векторе  $\hat{x}$ , разрешающем треугольную систему  $Rx = c$ .

При фиксированной точности арифметики MGS обладает примерно той же численной устойчивостью, что и алгоритм Хаусхолдера. Анализ ошибок, проведенный в работе [23], дает следующий результат: решение задачи  $Ax \cong b$ , вычисленное посредством MGS в арифметике со смешанной точностью  $(\eta, \eta^2)$ , является точным решением возмущенной задачи  $(A + E)x \cong b + f$ , где

$$\|E\|_F \leq 2n^{3/2} \|A\|_F \eta, \quad (19.35)$$

$$\|f\| \leq 2n^{3/2} \|b\| \eta. \quad (19.36)$$

Проведенное тестирование машинных программ (см. [185]) показало, что программы методов Хаусхолдера и MGS дают, по существу, одинаковую точность.

Число арифметических операций в MGS несколько больше, чем в методе Хаусхолдера (см. табл. 19.1), по той причине, что в MGS все операции производятся над векторами длины  $m$ , в то время как в методе Хаусхолдера столбцы последовательно укорачиваются. По той же причине программы MGS требуют обычно больше памяти, чем программы Хаусхолдера: в MGS нет удобного способа получать матрицу  $R$  на том месте, где хранилась исходная матрица  $A$ . Если, однако, не сохранять векторы  $q_i$ , то возможна такая организация MGS, что  $R$  замещает  $A$  в памяти; это достигает-

ся за счет некоторой дополнительной работы по перемещению хранимой информации.

Метод MGS можно приспособить и к тому, чтобы последовательно накапливать строки или группы строк матрицы  $[A : b]$ ; это бывает нужно в случае, когда произведение  $mn$  очень велико и  $m \gg n$ . Возможность такой организации основывается на том, что  $(n+1) \times (n+1)$ -матрица  $\tilde{D}\tilde{R}$  (см. (19.34)) определяет ту же задачу наименьших квадратов, что и  $m \times (n+1)$ -матрица  $[A : b]$ . Исходя из этого, можно построить последовательный алгоритм MGS точно таким же образом, каким соответствующее свойство триангуляризации Хаусхолдера используется в гл. 27 для построения последовательного метода Хаусхолдера.

Специальными адаптациями процесса Грама–Шмидта являются: 1) метод сопряженных градиентов, предназначенный для решения положительно определенных систем линейных уравнений (см. [94, 110, 151, с. 62–67] и имеющий некоторые качества, привлекательные в случае больших разреженных систем [154–156]; 2) метод Форсайта [55] для полиномиального сглаживания с использованием ортогонализированных многочленов.

#### У п р а ж н е н и я

19.37. Пусть  $R$  – верхняя треугольная  $n \times n$ -матрица, у которой некоторые диагональные элементы  $r_{ii}$  равны нулю. Показать, что существует верхняя треугольная  $n \times n$ -матрица  $U$  такая, что  $U^T U = R^T R$  и  $u_{ij} = 0, j = i, \dots, n$ , если  $r_{ii} = 0$ .

Указание. Построить  $U$  в виде  $U = QR$ , где  $Q$  – произведение конечного числа специальных образом подобранных вращений Гивенса.

19.38 [142]. Предположим, что  $\text{rank } A_{m \times n} = n$  и  $m > n$ . Посредством гауссова исключения (с частичным выбором главного элемента) можно получить разложение  $A = PLR$ , где  $L_{m \times n}$  – нижняя треугольная матрица с единичными диагональными элементами,  $R_{n \times n}$  – верхняя треугольная, а  $P$  – матрица перестановки, учитывающая перемещения строк.

а) Подсчитать количество операций при вычислении этого разложения.

б) Задачу НК можно решить, применяя метод Холесского к системе  $L^T L y = L^T P^T b$ , а затем определяя  $x$  из системы  $Rx = y$ . Подсчитать количество операций при формировании  $L^T L$  и при вычислении факторизации Холесского этой матрицы.

с) Показать, что общее количество операций в этом методе такое же, что и в методе Хаусхолдера (см. табл. 19.1). (Достаточно определить лишь коэффициенты при  $m^2$  и  $n^3$  в формулах для числа операций.)

19.39. Пусть  $A$  –  $m \times n$ -матрица ранга  $n$ , и пусть  $A = QR$  – разложение матрицы  $A$ , полученное применением к ней (в точной арифметике) модифицированного алгоритма Грама–Шмидта. Предположим, что  $Q$  и  $R$  нормированы так, чтобы столбцы  $Q$  имели единичную евклидову длину. Пусть

$$\begin{bmatrix} B_{n \times n} \\ C_{m \times n} \end{bmatrix}$$

есть матрица, полученная (в точной арифметике) применением к  $(m+n) \times n$ -матрице

$$\begin{bmatrix} 0_{n \times n} \\ A_{m \times n} \end{bmatrix}$$

алгоритма НФТ (см. 11.4). Показать, что  $B$  совпадает с  $R$  с точностью до знаков строк, а  $C$  совпадает с  $Q$  с точностью до знаков столбцов \*).

\*) Нужно учесть только, что столбцы  $C$  не нормированы. (Примеч. пер.)

19.40 (метод Холесского без квадратных корней). Вместо (19.5) рассмотрим разложение  $W^T DW = P$ , где  $W$  – верхняя треугольная с единичными диагональными элементами матрица, а  $D$  – диагональная матрица. Вывести формулы, аналогичные (19.12)–(19.14), для вычисления диагональных элементов  $D$  и наддиагональных элементов  $W$ . Показать, что, используя такое разложение матрицы  $\tilde{P}$  из (19.8), задачу (19.1) можно решить без вычисления квадратных корней.

## ГЛАВА 20

### ЛИНЕЙНЫЕ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ С ЛИНЕЙНЫМИ ОГРАНИЧЕНИЯМИ-РАВЕНСТВАМИ: РЕШЕНИЕ С ПОМОЩЬЮ БАЗИСА НУЛЬ-ПРОСТРАНСТВА

В этой главе мы начинаем изучение задач наименьших квадратов, в которых переменные должны удовлетворять некоторым ограничениям в форме линейных уравнений или неравенств. Такие задачи возникают во многих приложениях. Например, при вычерчивании кривой по точкам ограничения-равенства могут возникать из необходимости интерполировать часть данных или из требования непрерывности кривой, а может быть, и ее производных в некоторых точках. Ограничения-неравенства возникают из таких условий, как положительность, монотонность и выпуклость.

Мы будем кратко обозначать линейную задачу наименьших квадратов с линейными ограничениями-уравнениями как задачу НКУ. Задачу с линейными ограничениями-неравенствами (в которой могут быть и ограничения-уравнения) будем называть задачей НКН. В этой и двух последующих главах изложены три различных алгоритма решения задачи НКУ. Задача НКН будет рассмотрена в гл. 23.

Чтобы зафиксировать обозначения, дадим развернутую формулировку задачи НКУ.

**З а д а ч а НКУ 20.1** (наименьшие квадраты с ограничениями-уравнениями). Пусть даны  $m_1 \times n$ -матрица  $C$  ранга  $k_1$ ,  $m_1$ -вектор  $d$ ,  $m_2 \times n$ -матрица  $E$  и  $m_2$ -вектор  $f$ . Среди всех  $n$ -векторов  $x$ , удовлетворяющих системе  $Cx = d$ ,

$$(20.2)$$

найти вектор, который минимизирует величину

$$\|Ex - f\|. \quad (20.3)$$

Очевидно, что задача НКУ имеет решение тогда и только тогда, когда система (20.2) совместна. Мы будем предполагать совместность (20.2) на протяжении гл. 20–23. В обычных для практики вариантах этой задачи выполняются условия  $n > m_1 = k_1$ , обеспечивающие совместность системы (20.2) и существование более чем одного решения.

В дальнейшем будет показано, что если решение задачи НКУ существует, то оно единственно тогда и только тогда, когда расширенная матрица

$\begin{bmatrix} C \\ E \end{bmatrix}$  имеет ранг  $n$ . В случае неединственности имеется единственное решение минимальной длины.



Ясно, что задачу НКУ можно обобщить на тот случай, когда система (20.2) несовместна и интерпретируется в смысле наименьших квадратов. Нам не встречались практические ситуации такого рода; однако наше обсуждение задачи НКУ переносится с точностью до небольших модификаций и на этот случай.

Все три алгоритма для задачи НКУ компактны в том смысле, что не требуются двумерных массивов машинной памяти, за исключением тех, что хранят исходные данные задачи. Каждый из методов можно разделить на три этапа:

1. Сведение исходной задачи к безусловной задаче наименьших квадратов меньшей размерности.
2. Решение редуцированной задачи.
3. Преобразование решения редуцированной задачи в решение исходной задачи с ограничениями.

В первом методе [90], который будет описан в этой главе, используется ортогональный базис ядра матрицы ограничений. К исходной матрице  $\begin{bmatrix} C \\ E \end{bmatrix}$  применяются левые и правые ортогональные преобразования.

По отношению к задачам с неединственным решением метод обладает таким свойством: (единственное) нормальное решение редуцированной безусловной задачи порождает (единственное) решение минимальной длины для исходной задачи с ограничениями. Поэтому метод рекомендуется в ситуации, когда задача вполне может иметь недостаточный ранг, и приходится использовать методы стабилизации (см. гл. 25), основанные на идее ограничения длины вектора решения.

Вместе с численно устойчивыми приемами модификации метод может быть применен к решению последовательности задач НКУ, в которой очередная матрица  $C$  получается из предыдущей добавлением или удалением строк. Именно таким образом метод использован в [174] для построения алгоритма, решающего задачу НКН.

Кроме всего прочего метод представляет и теоретический интерес. С его помощью в теореме 20.31 показано существование безусловной задачи наименьших квадратов, эквивалентной задаче НКУ в том смысле, что обе имеют одно и то же множество решений для любых правых частей  $d$  и  $f$ . Это позволяет применять к задаче НКУ такие численные процедуры, как, например, сингулярное разложение.

Второй метод, описываемый в гл. 21, основан на прямом исключении и использует как ортогональные, так и неортогональные левые преобразования.

Каждый из этих двух методов можно использовать для удаления ограничений-равенств (если таковые имеются) в качестве первого шага в решении задачи НКН. Кроме того, оба метода адаптируются к последовательной обработке данных, когда после триангуляризации исходной матрицы добавляются новые строки к системе ограничений  $[C:d]$  или системе наименьших квадратов  $[E:f]$ .

Третий метод для задачи НКУ, изложенный в гл. 22, освещает некоторые важные свойства преобразования Хаусхолдера в применении к задаче наи-

меньших квадратов с сильно различающимися весами. С практической точки зрения главное достоинство этого метода состоит в следующем: он дает возможность решать задачу НКУ в случае, когда пользователь имеет доступ к подпрограмме метода Хаусхолдера для безусловной задачи наименьших квадратов, но не имеет и не хочет составлять сам программу одного из алгоритмов, специально рассчитанных на задачу НКУ.

Перейдем теперь к описанию нашего первого метода для задачи НКУ. В методе используется явное параметрическое представление элементов линейного многообразия

$$X = \{x: Cx = d\}. \quad (20.4)$$

Если

$$C = HRK^T \quad (20.5)$$

есть произвольное ортогональное разложение (определение см. в гл. 2) матрицы  $C$ , а  $K$  представлена в виде

$$K = \begin{bmatrix} K_1 & K_2 \end{bmatrix} n, \quad (20.6)$$

$\underbrace{\quad}_{k_1} \quad \underbrace{\quad}_{n-k_1}$

то (см. теорему 2.3, равенство (2.8))  $X$  можно представить как

$$X = \{x: x = \bar{x} + K_2 y_2\}, \quad (20.7)$$

где

$$\bar{x} = C^* d, \quad (20.8)$$

а  $y_2$  пробегает пространство всех векторов размерности  $n - k_1$ .

**Теорема 20.9.** Если система (20.2) совместна, то задача НКУ имеет единственное решение минимальной длины, задаваемое формулой

$$\hat{x} = C^* d + (EZ)^*(f - EC^* d), \quad (20.10)$$

где

$$Z = I_n - C^* C,$$

или, что эквивалентно,

$$\hat{x} = C^* d + K_2 (EK_2)^*(f - EC^* d). \quad (20.11)$$

Матрица  $K_2$  определена в (20.6).

Вектор  $\hat{x}$  является единственным решением задачи НКУ тогда и только

тогда, когда система (20.2) совместна и ранг матрицы  $\begin{bmatrix} C \\ E \end{bmatrix}$  равен  $n$ .

**Доказательство.** Установим эквивалентность выражений (20.10) и (20.11). Заметим прежде всего, что  $Z = I_n - C^* C = KK^T - KR^* RK^T =$

$$= KK^T - K \begin{bmatrix} I_{k_1} & 0 \\ 0 & 0 \end{bmatrix} K^T = KK^T - K_1 K_1^T = K_2 K_2^T. \text{ Поэтому } EZ = EK_2 K_2^T.$$

Равенство  $(EK_2 K_2^T)^* = K_2 (EK_2)^*$  можно проверить прямой подстановкой в условия Пенроуза (теорема 7.9), учитывая, что  $K_2^T K_2 = I_{n-k_1}$ . Таким образом, (20.10) и (20.11) эквивалентны.

Представление (20.7) показывает, что задача минимизации величины (20.3) по всем  $x \in X$  эквивалентна отысканию такого  $n - k_1$ -вектора  $y_2$ , который минимизирует величину  $\|E(\bar{x} + K_2 y_2) - f\|$ , или, что то же самое,

$$\|(EK_2)y_2 - (f - E\bar{x})\|. \quad (20.12)$$

Согласно (7.5), единственное нормальное решение этой задачи выражается формулой

$$\hat{y}_2 = (EK_2^T)^*(f - E\bar{x}). \quad (20.13)$$

В соответствии с (20.7) решение задачи НКУ запишется как

$$\hat{x} = \bar{x} + K_2 \hat{y}_2, \quad (20.14)$$

что эквивалентно (20.11), а также (20.10).

Поскольку столбцы  $K_2$  попарно ортогональны \*) и все вместе ортогональны к  $\bar{x}$ , то для нормы любого вектора  $x \in X$

$$\|x\|^2 = \|\bar{x}\|^2 + \|y_2\|^2. \quad (20.15)$$

Если вектор  $\tilde{y}_2 \neq \hat{y}_2$  также минимизирует (20.12), то  $\|\tilde{y}_2\| > \|\hat{y}_2\|$ . Следовательно,  $\tilde{x} = \bar{x} + K_2 \tilde{y}_2$  является решением задачи НКУ; при этом

$$\|\tilde{x}\|^2 = \|\bar{x}\|^2 + \|\tilde{y}_2\|^2 > \|\bar{x}\|^2 + \|\hat{y}_2\|^2 = \|\hat{x}\|^2. \quad (20.16)$$

Тем самым  $\hat{x}$  — единственное решение минимальной длины для задачи НКУ.

Остается связать единственность решения задачи НКУ со значением  $k$ :

$$k = \text{rang} \begin{bmatrix} C \\ E \end{bmatrix}.$$

Ясно, что при  $k < n$  существует  $n$ -вектор  $w \neq 0$ , для которого

$$\begin{bmatrix} C \\ E \end{bmatrix} w = 0, \quad (20.17)$$

и если  $\hat{x}$  — решение задачи НКУ, то  $\hat{x} + w$  — также решение.

Пусть, с другой стороны,  $k = n$ . Рассмотрим  $(m_1 + m_2) \times n$ -матрицу

$$M = \begin{bmatrix} C \\ E \end{bmatrix} K = \begin{bmatrix} CK_1 & CK_2 \\ EK_1 & EK_2 \end{bmatrix} \begin{matrix} \} m_1 \\ \} m_2 \end{matrix}, \quad (20.18)$$

$$\begin{matrix} \underbrace{\quad}_{k_1} & \underbrace{\quad}_{n-k_1} \end{matrix}$$

ранг которой также равен  $n$ . Так как у этой матрицы  $n$  столбцов, они

\*) И нормированы. (Примеч. пер.)

должны быть линейно независимыми. Поскольку  $CK_2 = 0$ ,  $\text{rang } EK_2$  непременно равен  $n - k_1$ . Поэтому  $\hat{y}$  из (20.13) — единственный вектор, минимизирующий (20.12), а  $\hat{x}$  из (20.14) — единственное решение задачи НКУ. Теорема 20.9 доказана.

Вычислительную процедуру можно построить на базе формулы (20.11). Будем считать, что  $k_1 = m_1$ , так как это — обычный для практики случай. Тогда в качестве матрицы  $H$  в (20.5) можно взять единичную матрицу.

Пусть  $K$  — ортогональная  $n \times n$ -матрица, которая, будучи умножена на  $C$  слева, преобразует  $C$  к нижней треугольной матрице. Умножим  $C$  и  $E$  справа на  $K$ ; тогда

$$\begin{bmatrix} C \\ E \end{bmatrix} K = \begin{bmatrix} \tilde{C}_1 & 0 \\ \tilde{E}_1 & \tilde{E}_2 \end{bmatrix} \begin{matrix} \} m_1 \\ \} m_2 \end{matrix}, \quad (20.19)$$

$$\begin{matrix} \underbrace{\quad}_{m_1} & \underbrace{\quad}_{n-m_1} \end{matrix}$$

где  $\tilde{C}_1$  — нижняя треугольная невырожденная  $m_1 \times m_1$ -матрица.

Найдем  $m_1$ -вектор  $\hat{y}_1$  из нижней треугольной системы

$$\tilde{C}_1 y_1 = d. \quad (20.20)$$

Вычислим вектор

$$\tilde{f} = f - \tilde{E}_1 \hat{y}_1. \quad (20.21)$$

Решим задачу наименьших квадратов

$$\tilde{E}_2 y_2 \cong \tilde{f} \quad (20.22)$$

относительно  $n - m_1$ -вектора  $\hat{y}_2$ . Наконец, вычислим решение исходной задачи

$$x = K \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}. \quad (20.23)$$

Шаги реального вычислительного процесса описываются нижеследующим алгоритмом. Поначалу матрицы и векторы  $C, d, E$  и  $f$  хранятся в массивах с теми же именами. Массив  $g$  имеет длину  $m_1$ . Каждый из массивов  $h, u$  и  $p$  имеет длину  $n - m_1$ . Решение минимальной длины будет помещено в массив  $x$  длины  $n$ . Параметр  $\tau$  — это задаваемый пользователем допуск, нужный на шаге 6.

Алгоритм 20.24. LSE( $C, d, E, f, m_1, m_2, n, g, h, u, p, x, \tau$ ):

1. Для  $i := 1, \dots, m_1$  выполнить алгоритм Н1( $i, i+1, n, c_{i1}, g_i, c_{i+1,1}, m_1 - i$ ) (см. (20.19)). Здесь и на шаге 2 алгоритмы Н1 и Н2 оперируют со строками массивов  $C$  и  $E$ .

2. Для  $i := 1, \dots, m_1$  выполнить алгоритм Н2( $i, i+1, n, c_{i1}, g_i, e_{i1}, m_2$ ) (см. (20.19)). Если матрицы  $C$  и  $E$  хранятся единым  $(m_1 + m_2) \times n$ -массивом (назовем его  $W$ ), то шаги 1 и 2 можно объединить: для  $i := 1, \dots, m_1$  выполнить алгоритм Н1( $i, i+1, n, w_{i1}, g_i, w_{i+1,1}, m_1 + m_2 - i$ ).

3. Положить  $x_1 := d_1/c_{11}$ . Если  $m_1 = 1$ , перейти к шагу 5.

4. Для  $i := 2, \dots, m_1$  положить (см. (20.20))

$$x_i := \frac{1}{c_{ii}} \left( d_i - \sum_{j=1}^{i-1} c_{ij} x_j \right).$$

5. Для  $i := 1, \dots, m_2$  положить (см. (20.21))

$$f_i := f_i - \sum_{j=1}^{m_1} e_{ij} x_j.$$

6. Выполнить в соответствии с (14.9) алгоритм HFTI( $e_{1,m_1+1}, m_2, n - m_1, f, \tau, x_{m_1+1}, k, h, u, p$ ). (Решение  $\hat{y}_2$  задачи (20.22) теперь вычислено и помещено в ячейки  $x_i, i := m_1 + 1, \dots, n$ .)

7. Для  $i := m_1, m_1 - 1, \dots, 1$  выполнить алгоритм H2( $i, i + 1, n, c_{ii}, g_i, x, 1$ ) (см. (20.23)). Отметим, что в списке параметров алгоритма H2 содержится ссылка на строки массива  $C$ , а преобразования применяются к одномерному массиву  $x$ .

8. *Комментарий.* В массиве  $x$  теперь находится решение минимальной длины задачи НКУ.

В качестве примера задачи НКУ рассмотрим минимизацию  $\|Ex - f\|$  при условии  $Sx = d$ , где

$$C = \begin{bmatrix} 0,4087 & 0,1593 \end{bmatrix}, \quad (20.25)$$

$$E = \begin{bmatrix} 0,4302 & 0,3516 \\ 0,6246 & 0,3384 \end{bmatrix}, \quad (20.26)$$

$$d = 0,1376, \quad (20.27)$$

$$f = \begin{bmatrix} 0,6593 \\ 0,9666 \end{bmatrix}. \quad (20.28)$$

Ортогональная матрица Хаусхолдера, которая приводит  $C$  к треугольному виду, здесь такова:

$$K = \begin{bmatrix} -0,9317 & -0,3632 \\ -0,3632 & 0,9317 \end{bmatrix}.$$

Конечно, при выполнении алгоритма LSE (см. (20.24)) эту матрицу обычно не вычисляют в явном виде. Теперь в соответствии с формулами (20.19) – (20.23) получаем

$$\begin{bmatrix} \tilde{C}_1 & 0 \\ \tilde{E}_1 & \tilde{E}_2 \end{bmatrix} = \begin{bmatrix} -0,4386 & 0,0 \\ -0,5285 & 0,1714 \\ -0,7049 & 0,0885 \end{bmatrix}.$$

$$\hat{y}_1 = \frac{0,1376}{-0,4386} = -0,3137, \quad \tilde{f} = \begin{bmatrix} 0,4935 \\ 0,7455 \end{bmatrix}, \quad \hat{y}_2 = 4,0472,$$

$$\hat{x} = \begin{bmatrix} -1,1775 \\ 3,8848 \end{bmatrix}. \quad (20.29)$$

Рассмотрим теперь некоторые теоретические следствия теоремы 20.9.

Формулу (20.10) можно записать в виде  $\hat{x} = \hat{A}^* b$ , если положить  $b = \begin{bmatrix} d \\ f \end{bmatrix}$  и

$$\hat{A}^* = [C^* - (EZ)^* E C^* : (EZ)^*] = [C^* - K_2 (E K_2)^* E C^* : K_2 (E K_2)^*]. \quad (20.30)$$

Но (20.30) означает, что  $\hat{x}$  есть нормальное псевдорешение задачи  $\min_x \|\hat{A}x - b\|$ , где  $\hat{A}$  – матрица, псевдообратная для матрицы  $\hat{A}^*$ , определенной в (20.30).

**Т е о р е м а 20.31.** Псевдообратная для матрицы  $\hat{A}^*$ , определенной в (20.30), имеет вид

$$\hat{A} = \begin{bmatrix} C \\ \hat{E} \end{bmatrix}, \quad (20.32)$$

где

$$\hat{E} = (EZ)(EZ)^* E = E(EZ)^* E = E K_2 (E K_2)^* E, \quad (20.33)$$

а  $Z$  и  $K_2$  определены так же, как в теореме 20.9.

*Доказательство.* Напомним, что

$$C K_2 = 0, \quad K_2^T C^* = 0,$$

$$K_2^T K_2 = I_{n-m_1}, \quad Z = I_n - C^* C = K_2 K_2^T,$$

$$(EZ)^* = (E K_2 K_2^T)^* = K_2 (E K_2)^*.$$

Из этих соотношений прямо следуют другие, также используемые в доказательстве:

$$E Z C^* = E K_2 K_2^T C^* = 0,$$

$$C (EZ)^* = C K_2 (E K_2)^* = 0,$$

$$E (EZ)^* = E K_2 (E K_2)^* =$$

$$= E K_2 K_2^T K_2 (E K_2)^* = (EZ)(EZ)^*. \quad (20.34)$$

Равенство (20.34) устанавливает тождество различных выражений для  $\hat{E}$  в (20.33).

Положим теперь

$$X = \begin{bmatrix} C \\ (EZ)(EZ)^*E \end{bmatrix}$$

и проверим четыре условия Пенроуза (теорема 7.9).

$$\begin{aligned} 1. \hat{A}^*X &= C^*C - (EZ)^*EC^*C + (EZ)^*(EZ)(EZ)^*E = \\ &= C^*C - (EZ)^*E(C^*C - I) = C^*C + (EZ)^*(EZ). \end{aligned}$$

Полученная матрица симметрична.

$$\begin{aligned} 2. X\hat{A}^* &= \begin{bmatrix} CC^* - C(EZ)^*EC^* & C(EZ)^* \\ (EZ)(EZ)^*E[I - (EZ)^*E]C^* & (EZ)(EZ)^*E(EZ)^* \end{bmatrix} = \\ &= \begin{bmatrix} CC^* & 0 \\ 0 & (EZ)(EZ)^* \end{bmatrix}. \end{aligned}$$

Полученная матрица симметрична.

$$\begin{aligned} 3. \hat{A}^*X\hat{A}^* &= [C^*C + (EZ)^*(EZ)][C^* - (EZ)^*EC^* : (EZ)^*] = \\ &= \{C^* - (EZ)^*(EZ)[I - (EZ)^*E]C^* : (EZ)^*(EZ)(EZ)^*\} = \\ &= [C^* - (EZ)^*(EZ)(EZ)^*EC^* : (EZ)^*] = \\ &= [C^* - (EZ)^*EC^* : (EZ)^*] = \hat{A}^*. \end{aligned}$$

$$\begin{aligned} 4. X\hat{A}^*X &= \begin{bmatrix} C \\ (EZ)(EZ)^*E \end{bmatrix} \cdot [C^*C + (EZ)^*EZ] = \\ &= \begin{bmatrix} C \\ (EZ)(EZ)^*EC^*C + (EZ)(EZ)^*E(EZ)^*EZ \end{bmatrix} = \\ &= \begin{bmatrix} C \\ (EZ)(EZ)^*E[C^*C + Z] \end{bmatrix} = \begin{bmatrix} C \\ (EZ)(EZ)^*E \end{bmatrix} = X. \end{aligned}$$

Теорема 20.31 доказана.

В качестве числовой иллюстрации к теореме 20.31 вычислим матрицы (20.32) и (20.33) для задачи (20.25) – (20.28). Из (20.33) имеем

$$\begin{aligned} \hat{E} &= (EK_2)(EK_2)^*E = \begin{bmatrix} 0,1714 \\ 0,0885 \end{bmatrix} \cdot [4,6076 \quad 2,3787] \times \\ &\times \begin{bmatrix} 0,4302 & 0,3516 \\ 0,6246 & 0,3384 \end{bmatrix} = \begin{bmatrix} 0,5943 & 0,4155 \\ 0,3068 & 0,2145 \end{bmatrix}. \end{aligned}$$

а из (20.32) находим

$$\hat{A} = \begin{bmatrix} C \\ \hat{E} \end{bmatrix} = \begin{bmatrix} 0,4087 & 0,1593 \\ 0,5943 & 0,4155 \\ 0,3068 & 0,2145 \end{bmatrix}.$$

Согласно теореме 20.31, полученная матрица обладает следующим свойством. Для произвольного одномерного вектора  $\bar{d}$  и произвольного двумерного вектора  $\bar{f}$  решение задачи наименьших квадратов  $\hat{A}x \cong \begin{bmatrix} \bar{d} \\ \bar{f} \end{bmatrix}$  является

в то же время решением такой задачи НКУ: минимизировать  $\|Ex - \bar{f}\|$  при условии  $Cx = \bar{d}$ , где  $C$  и  $E$  указаны в (20.25) и (20.26). Нужно отметить, однако, что нормы невязок для этих двух задач наименьших квадратов в общем случае неодинаковы.

## ГЛАВА 21

### ЛИНЕЙНЫЕ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ С ЛИНЕЙНЫМИ ОГРАНИЧЕНИЯМИ-РАВЕНСТВАМИ: РЕШЕНИЕ ПОСРЕДСТВОМ ПРЯМОГО ИСКЛЮЧЕНИЯ

К введенной в предыдущей главе задаче НКУ здесь будет применен метод прямого исключения. Будем считать, что ранг  $k_1$  матрицы  $C$  равен  $m_1$ , а ранг матрицы  $\begin{bmatrix} C \\ E \end{bmatrix}$  равен  $n$ . Эти предположения гарантируют, согласно теореме 20.9, что задача НКУ имеет единственное решение для любых правых частей  $d$  и  $f$ .

Мы предположим еще, что столбцы матрицы  $\begin{bmatrix} C \\ E \end{bmatrix}$  упорядочены так, чтобы первые  $m_1$  столбцов  $C$  были линейно независимы. Перестановки столбцов, обеспечивающие такое упорядочение, составляют обязательную часть любой вычислительной процедуры того типа, какой обсуждается в данной главе.

Вспользуемся представлениями

$$\begin{bmatrix} C \\ E \end{bmatrix} = \begin{bmatrix} C_1 & C_2 \\ E_1 & E_2 \end{bmatrix} \begin{matrix} \} m_1 \\ \} m_2 \end{matrix}, \quad (21.1)$$

$\underbrace{\hspace{1.5cm}}_{m_1} \quad \underbrace{\hspace{1.5cm}}_{n-m_1}$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{matrix} \} m_1 \\ \} n - m_1 \end{matrix}. \quad (21.2)$$

Систему ограничений  $Cx = d$  можно разрешить относительно  $x_1$ :

$$x_1 = C_1^{-1}(d - C_2 x_2).$$

Подставляя это выражение для  $x_1$  в  $\|Ex - f\|$ , получаем

$$\begin{aligned} \|Ex - f\| &= \|E_1 C_1^{-1}(d - C_2 x_2) + E_2 x_2 - f\| = \\ &= \|(E_2 - E_1 C_1^{-1} C_2) x_2 - (f - E_1 C_1^{-1} d)\| \equiv \|\tilde{E}_2 x_2 - \tilde{f}\|. \end{aligned} \quad (21.3)$$

Вектор  $x_2$  определяется из условия минимума этой величины.

С принципиальной точки зрения процедура решения состоит из следующих этапов. Вычисляем матрицу

$$\tilde{E}_2 = E_2 - E_1 C_1^{-1} C_2 \quad (21.4)$$

и вектор

$$\tilde{f} = f - E_1 C_1^{-1} d. \quad (21.5)$$

Решаем задачу наименьших квадратов

$$\tilde{E}_2 x_2 \cong \tilde{f} \quad (21.6)$$

и, наконец, вычисляем

$$x_1 = C_1^{-1}(d - C_2 x_2). \quad (21.7)$$

Эти этапы можно организовать многими различными способами. Например [25], предположим, что вычислено  $QR$ -разложение матрицы  $C_1$ :

$$C_1 = Q_1^T \tilde{C}_1, \quad (21.8)$$

где  $Q_1$  — ортогональная матрица, а  $\tilde{C}_1$  — правая треугольная матрица. Тогда равенства (21.4), (21.5) можно записать в виде

$$\tilde{E}_2 = E_2 - (E_1 \tilde{C}_1^{-1}) Q_1 C_2 \equiv E_2 - \tilde{E}_1 \tilde{C}_2, \quad (21.9)$$

$$\tilde{f} = f - (E_1 \tilde{C}_1^{-1}) Q_1 d \equiv f - \tilde{E}_1 \tilde{d}. \quad (21.10)$$

В результате получается следующий алгоритм. Вычисляются преобразования Хаусхолдера, которые приводят  $C_1$  к треугольному виду; эти же преобразования применяются к  $C_2$  и  $d$ :

$$Q_1 [C_1 : C_2 : d] = [\tilde{C}_1 : \tilde{C}_2 : \tilde{d}]. \quad (21.11)$$

Далее вычисляют  $m_2 \times m_1$ -матрицу  $\tilde{E}_1$  как решение треугольных систем

$$\tilde{E}_1 \tilde{C}_1 = E_1. \quad (21.12)$$

Затем вычисляют матрицу

$$\tilde{E}_2 = E_2 - \tilde{E}_1 \tilde{C}_2 \quad (21.13)$$

и вектор

$$\tilde{f} = f - \tilde{E}_1 \tilde{d}. \quad (21.14)$$

Потом матрицу  $\tilde{E}_2$  посредством преобразований Хаусхолдера приводят

к треугольному виду; эти же преобразования применяются к  $\tilde{f}$ :

$$Q_2 [\tilde{E}_2 : \tilde{f}] = \begin{bmatrix} \tilde{E}_2 \hat{f} \\ 0 \quad \varphi \\ 0 \quad 0 \end{bmatrix} \begin{matrix} \} n - m_1 \\ \} 1 \\ \} m_1 + m_2 - n - 1 \end{matrix} \quad (21.15)$$

Наконец, из треугольной системы

$$\begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \\ 0 & \tilde{E}_2 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \tilde{d} \\ \tilde{f} \end{bmatrix} \quad (21.16)$$

определяется решение

$$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}.$$

В описанном алгоритме не требуется двумерных массивов машинной памяти, за исключением тех, что нужны для хранения исходных данных задачи. При этом величины, помеченные знаком  $\sim$ , следует записывать на место одноименных непомеченных величин, а величины со знаком  $\wedge$  на место одноименных величин со знаком  $\sim$ .

Подпрограмма этого алгоритма может быть очень компактной, как показывает алгоритмическая процедура decompose, приведенная в работе [25]

Предположим, что матрица исходных данных  $\begin{bmatrix} C \\ E \end{bmatrix}$  хранится в  $m \times n$ -массиве  $A$ , где  $m = m_1 + m_2$ , а вектор исходных данных  $\begin{bmatrix} d \\ f \end{bmatrix}$  хранится в массиве  $b$  длины  $m$ . Шаги (21.11) и (21.15) может выполнять одна и та же подпрограмма, реализующая, по существу, шаги 2–12 алгоритма HFTI (см. 14.9).

На этапе (21.11) арифметические операции выполняются лишь с первыми  $m_1$  строками матрицы  $[A:b]$ , но любые необходимые перестановки столбцов должны производиться над полными  $m$ -мерными столбцами  $A$ . Шаги (21.12)–(21.14), которые можно интерпретировать как гауссовы исключения, реализуются следующими операциями:

$$a_{i1} = \frac{a_{i1}}{a_{11}}, \quad (21.17)$$

$$a_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} a_{ik} a_{kj}}{a_{ij}}, \quad j = 2, \dots, m_1, \quad (21.18)$$

$$a_{ij} = a_{ij} - \sum_{k=1}^{m_1} a_{ik} a_{kj}, \quad j = m_1 + 1, \dots, n, \quad (21.19)$$

$$b_i = b_i - \sum_{k=1}^{m_1} a_{ik} b_k, \quad i = m_1 + 1, \dots, m. \quad (21.20)$$

В качестве примера для этой процедуры снова рассмотрим задачу НКУ с исходными данными (20.25)–(20.28). Для этой задачи матрицу  $Q_1$

в (21.11) можно считать единичной матрицей порядка 1. В соответствии с формулами (21.11) – (21.16) вычисляем

$$\tilde{E}_2 = \begin{bmatrix} 0,1839 \\ 0,0949 \end{bmatrix}, \quad \tilde{f} = \begin{bmatrix} 0,5145 \\ 0,7563 \end{bmatrix}, \quad \hat{x} = \begin{bmatrix} -1,1775 \\ 3,8848 \end{bmatrix}.$$

## ГЛАВА 22

### ЛИНЕЙНЫЕ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ С ЛИНЕЙНЫМИ ОГРАНИЧЕНИЯМИ-РАВЕНСТВАМИ: РЕШЕНИЕ ПУТЕМ ВЗВЕШИВАНИЯ

Многие статистики и инженеры пользуются следующим эвристическим приемом. Пусть нужно решить линейную задачу наименьших квадратов, в которой желательно удовлетворить точно некоторые из уравнений. Этого можно добиться приближенно, приписывая большие веса соответствующим уравнениям и решая полученную задачу наименьших квадратов. К тому же результату можно прийти, если приписать малые веса прочим уравнениям и решить такую задачу наименьших квадратов.

В данной главе мы проведем анализ этой вычислительной процедуры решения задачи НКУ (20.1). Основная идея проста. Вычисляем решение задачи наименьших квадратов

$$\begin{bmatrix} C \\ \epsilon E \end{bmatrix} x \cong \begin{bmatrix} d \\ \epsilon f \end{bmatrix} \quad (22.1)$$

(используя, например, метод Хаусхолдера) для "малого", но ненулевого значения  $\epsilon$ . Тогда решение  $\tilde{x}(\epsilon)$  задачи (22.1) "близко" (в смысле, определенном соотношениями (22.38) и (22.37)) к решению  $\hat{x}$  задачи НКУ.

Такой общий подход представляет практический интерес по той причине, что некоторые существующие программы метода наименьших квадратов могут эффективно решать задачу НКУ путем решения задач вида (22.1).

Очевидным практическим изъяном этой идеи является то обстоятельство, что матрица задачи (22.1) становится как угодно плохо обусловленной (в предположении, что  $\text{rank } C < n$ ) по мере уменьшения параметра  $\epsilon$ . Это обстоятельство, конечно, ограничивает практическую применимость рассматриваемого подхода, если задача (22.1) решается методом нормальных уравнений (см. (19.1)). В самом деле, нормальные уравнения для (22.1) имеют вид

$$(C^T C + \epsilon^2 E^T E) x = C^T d + \epsilon^2 E^T f,$$

и если только участвующие здесь матрицы не обладают очень специальной структурой, то машинное представление матрицы  $C^T C + \epsilon^2 E^T E$  при достаточно малом  $\epsilon$  будет неотличимо от представления матрицы  $C^T C$  (а вектора  $C^T d + \epsilon^2 E^T f$  – от вектора  $C^T d$ ).

Тем не менее Пауэлл и Рид [146] обнаружили экспериментально, что приемлемое решение задачи с сильно различающимися весами типа задачи (22.1) можно получить преобразованиями Хаусхолдера, если позаботиться о введении в алгоритм специальных перестановок строк. Отсылаем читателя к гл. 17, где основные теоретические результаты работы [146] представлены в виде трех теорем. В названной работе построению  $k$ -го преобразования Хаусхолдера предшествуют следующие перестановки. Во-первых, как и в алгоритме НФТИ (см. 14.9), выполняется обычная перестановка столбцов, переводящая в положение  $k$ -го столбец с наибольшей евклидовой длиной отрезка от  $k$ -й строки до  $m$ -й. Затем просматриваются элементы столбца  $k$  с номерами  $k, \dots, m$ . Если элемент с наибольшим модулем находится в строке  $l$ , то строки  $l$  и  $k$  переставляются.

Для задачи (22.1) это правило перестановок будет большей частью чересчур перестраховочным. Главное соображение, лежащее в основе анализа Пауэлла и Рида, таково: если (в обозначениях формул (10.5) – (10.11)) главный элемент  $v_p$  значим для задачи, но незначителен по величине в сравнении с некоторыми элементами  $v_i$ ,  $i > p$ , то может произойти существенная потеря численной информации. Под "значим для задачи" мы подразумеваем, что ее решение сильно изменилось бы, если  $v_p$  заменить нулем. Но если  $|v_p|$  достаточно мал сравнительно с каким-либо числом  $|v_i|$ ,  $i > p$ , то при машинном вычислении  $s$  и  $u_p$  по формулам (10.5) и (10.7)  $u_p$  будет неотличим от точного нуля.

Предположим, что ненулевые элементы матриц  $C$  и  $E$  задачи (22.1) имеют одинаковый порядок, так что если и есть сильный разброс в порядках ненулевых коэффициентов матрицы этой задачи, то он связан только с малым параметром  $\epsilon$ . Тогда отмеченная выше катастрофическая потеря численной информации могла бы, как правило, произойти лишь в том случае, если бы главный элемент  $v_p$  (из (10.5)) был выбран из строки матрицы  $\epsilon E$  (скажем,  $v_p = \epsilon e_{ij}$ ), хотя некоторая строка  $C$  (скажем,  $l$ -я) еще не выбиралась в качестве ведущей и содержит такой элемент  $c_{lj}$ , что  $|c_{lj}| \gg \epsilon |e_{ij}|$ . Этой ситуации можно избежать, если сохранять порядок строк, указанный в (22.1), так, чтобы все строки  $C$  были использованы в качестве ведущих прежде, чем будет использована какая-либо строка  $\epsilon E$ . Здесь мы применили символы  $\epsilon E$  и  $C$  для обозначения как матриц исходной задачи, так и их наследниц в алгоритме.

Перейдем теперь к анализу сходимости решения задачи (22.1) к решению задачи НКУ при  $\epsilon \rightarrow 0$ . Рассмотрим вначале специальный случай задачи НКУ, когда  $C$  – диагональная матрица, а  $E$  – единичная матрица. Этот случай легко разобрать, а, как будет показано в дальнейшем, анализ более общих случаев сводится к этому специальному. Положим

$$C = \begin{bmatrix} s_1 & 0 & 0 & \dots & 0 \\ & \ddots & & & \\ 0 & & s_{m_1} & 0 & \dots & 0 \end{bmatrix}_{m_1 \times n}, \quad s_1 \geq \dots \geq s_{m_1} > 0, \quad (22.2)$$

$$E = I_n. \quad (22.3)$$

Для специального случая (22.2), (22.3) приводимые ниже две леммы показывают, что при  $\epsilon \rightarrow 0$  решение задачи наименьших квадратов (22.1) сходится к решению задачи НКУ.

**Лемма 22.4.** Пусть  $C$  и  $E$  – матрицы, определенные в (22.2) и (22.3). Тогда решение задачи НКУ в обозначениях (20.2), (20.3) запишется так:

$$\hat{x}_i = \begin{cases} \frac{d_i}{s_i}, & i = 1, \dots, m_1, \\ f_i, & i = m_1 + 1, \dots, n. \end{cases} \quad (22.5)$$

При этом длина невязки  $E\hat{x} - f$  равна

$$\|E\hat{x} - f\| = \left[ \left( f_1 - \frac{d_1}{s_1} \right)^2 + \dots + \left( f_{m_1} - \frac{d_{m_1}}{s_{m_1}} \right)^2 \right]^{1/2}. \quad (22.6)$$

**Лемма 22.7.** Пусть  $C$  и  $E$  – матрицы, определенные в (22.2) и (22.3). Единственное псевдорешение задачи (22.1) запишется так:

$$\tilde{x}_i = \tilde{x}_i(\epsilon) = \begin{cases} \frac{d_i}{s_i} + \epsilon^2 \left( f_i - \frac{d_i}{s_i} \right) \left\{ s_i^2 \left[ 1 + \left( \frac{\epsilon}{s_i} \right)^2 \right] \right\}^{-1}, & i = 1, \dots, m_1, \\ f_i, & i = m_1 + 1, \dots, n. \end{cases} \quad (22.8)$$

Лемму 22.7 легко проверить, построив и решив систему нормальных уравнений  $(C^T C + \epsilon^2 E^T E)x = C^T d + \epsilon^2 E^T f$ . При наших предположениях относительно  $C$  и  $E$  эта система имеет диагональную матрицу коэффициентов, откуда легко следует (22.8).

Чтобы удобно записать разность между  $\tilde{x}(\epsilon)$  и  $\hat{x}$ , определим векторно-значную функцию  $h(\epsilon)$  соотношением

$$\tilde{x}_i(\epsilon) - \hat{x}_i = \epsilon^2 h_i(\epsilon), \quad i = 1, \dots, n. \quad (22.9)$$

Согласно (22.5) и (22.8),

$$h_i(\epsilon) = \begin{cases} \frac{s_i^{-2} (f_i - d_i/s_i)}{1 + (\epsilon/s_i)^2}, & i = 1, \dots, m_1, \\ 0, & i = m_1 + 1, \dots, n, \end{cases} \quad (22.10)$$

$$|h_i(\epsilon)| \leq \begin{cases} s_i^{-2} \left| f_i - \frac{d_i}{s_i} \right| < s_{m_1}^{-2} \left| f_i - \frac{d_i}{s_i} \right| = s_{m_1}^{-2} \|f - \hat{x}\|, & i = 1, \dots, m_1, \\ 0, & i = m_1 + 1, \dots, n. \end{cases} \quad (22.11)$$

Вектор невязки для задачи (22.1) имеет вид

$$\begin{aligned} \begin{bmatrix} d \\ \epsilon f \end{bmatrix} - \begin{bmatrix} C \\ \epsilon E \end{bmatrix} \tilde{x}(\epsilon) &= \begin{bmatrix} d \\ \epsilon f \end{bmatrix} - \begin{bmatrix} C \\ \epsilon E \end{bmatrix} [\hat{x} + \epsilon^2 h(\epsilon)] = \\ &= \begin{bmatrix} -\epsilon^2 Ch(\epsilon) \\ \epsilon(f - E\hat{x}) - \epsilon^2 Eh(\epsilon) \end{bmatrix}. \end{aligned} \quad (22.12)$$

Введем взвешенную евклидову норму

$$\| \cdot \|_{\epsilon} = \left[ \underbrace{(\cdot)^2 + \dots + (\cdot)^2}_{m_1} + \underbrace{(\cdot/\epsilon)^2 + \dots + (\cdot/\epsilon)^2}_n \right]^{1/2}, \quad (22.13)$$

тогда

$$\left\| \begin{bmatrix} d \\ \epsilon f \end{bmatrix} - \begin{bmatrix} C \\ \epsilon E \end{bmatrix} \tilde{x}(\epsilon) \right\|_{\epsilon}^2 = \epsilon^4 \|Ch(\epsilon)\|^2 + \|f - E\hat{x} - \epsilon^2 Eh(\epsilon)\|^2. \quad (22.14)$$

Из (22.10) видим, что  $\tilde{x}(\epsilon) \rightarrow \hat{x}$  при  $\epsilon \rightarrow 0$ , а из (22.14) следует, что при  $\epsilon \rightarrow 0$

$$\left\| \begin{bmatrix} d \\ \epsilon f \end{bmatrix} - \begin{bmatrix} C \\ \epsilon E \end{bmatrix} \tilde{x}(\epsilon) \right\|_{\epsilon} \rightarrow \|f - E\hat{x}\|.$$

При практических вычислениях представляет интерес вопрос, насколько малым следует взять  $\epsilon$ , чтобы гарантировать неразличимость  $\tilde{x}(\epsilon)$  и  $\hat{x}$ , если их компоненты представлены машинными словами с относительной точностью  $\eta$ . Если все компоненты вектора  $\hat{x}$  ненулевые, то это условие обеспечивается требованием, чтобы

$$|\epsilon^2 h_i(\epsilon)| \leq \eta |\hat{x}_i|, \quad i = 1, \dots, n. \quad (22.15)$$

Оно будет удовлетворено при всех  $|\epsilon| \leq \epsilon_0$ , если

$$\epsilon_0^2 = \min_i \left\{ \frac{\eta s_i^2 |d_i/s_i|}{|f_i - d_i/s_i|} : f_i - d_i/s_i \neq 0 \right\}. \quad (22.16)$$

Согласно (22.9) и (22.11), для нормы разности между  $\tilde{x}(\epsilon)$  и  $\hat{x}$  можно дать оценку

$$\|\tilde{x}(\epsilon) - \hat{x}\| \leq \frac{\epsilon^2 \|f - \hat{x}\|}{s_{m_1}^2}. \quad (22.17)$$

Поэтому, чтобы обеспечить выполнение неравенства  $\|\tilde{x}(\epsilon) - \hat{x}\| \leq \eta \|\hat{x}\|$ , достаточно взять  $\epsilon$ , для которого  $|\epsilon| \leq \epsilon_0$ , где

$$\epsilon_0^2 = \frac{\eta s_{m_1}^2 \|\hat{x}\|}{\|f - \hat{x}\|}. \quad (22.18)$$

Заметим, что формула (22.18) неприменима, если  $\|f - \hat{x}\| = 0$ . В этом случае, однако, система (22.1) совместна и имеет одно и то же решение при всех  $\epsilon \neq 0$ .

Рассмотрим теперь задачу НКУ для более общего случая, когда  $C - m_1 \times n$ -матрица ранга  $k_1 \leq m_1 < n$ ,  $E - m_2 \times n$ -матрица и ранг матрицы  $[C^T : E^T]$  равен  $n$ . Кроме того, предполагается, что уравнения связей  $Cx = d$  совместны.

Подходящими заменами переменных мы сведем эту задачу к только что рассмотренному специальному случаю. Прежде всего заметим, что при  $|\epsilon| < 1$  задача (22.1) эквивалентна задаче наименьших квадратов

$$\begin{bmatrix} (1 - \epsilon^2)^{1/2} C \\ \epsilon E' \end{bmatrix} x \cong \begin{bmatrix} (1 - \epsilon^2)^{1/2} d \\ \epsilon f' \end{bmatrix}, \quad (22.19)$$

где

$$E' = \begin{bmatrix} C \\ E \end{bmatrix}, \quad f' = \begin{bmatrix} d \\ f \end{bmatrix}. \quad (22.20)$$

Теперь мы проведем равномерное масштабирование задачи (22.19), умножив коэффициенты обеих частей на  $(1 - \epsilon^2)^{-1/2}$ . В результате получим задачу

$$\begin{bmatrix} C \\ \rho E' \end{bmatrix} x \cong \begin{bmatrix} d \\ \rho f' \end{bmatrix}, \quad \rho = \epsilon(1 - \epsilon^2)^{-1/2}. \quad (22.21)$$

Рассмотрим  $QR$ -разложение матрицы  $E'$ :

$$E' = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix}. \quad (22.22)$$

Здесь  $Q$  — ортогональная матрица порядка  $m_1 + m_2$ , а  $R$  — невырожденная  $n \times n$ -матрица. Полагая

$$g = Qf', \quad Rx = y, \quad (22.23)$$

приходим к эквивалентной задаче наименьших квадратов

$$\begin{matrix} m_1 \{ \\ n \{ \\ m_1 + m_2 - n \{ \end{matrix} \begin{bmatrix} CR^{-1} \\ \rho I_n \\ 0 \end{bmatrix} y \cong \begin{bmatrix} d \\ \rho g \end{bmatrix} \begin{matrix} \} m_1 \\ \} m_1 + m_2 \\ \} m_1 + m_2 \end{matrix}. \quad (22.24)$$

Пусть

$$CR^{-1} = USV^T \quad (22.25)$$

есть сингулярное разложение матрицы  $CR^{-1}$  (см. теорему 4.1). Так как  $C$  имеет ранг  $k_1$ , то это же верно для  $CR^{-1}$ , и потому

$$S_{m_1 \times n} = \begin{bmatrix} s_1 & & & 0 \\ & \ddots & & \\ & & s_{k_1} & \\ 0 & & & 0 \end{bmatrix} \begin{matrix} \} k_1 \\ \\ \} m_1 - k_1 \end{matrix}$$

и  $s_1 \geq \dots \geq s_{k_1} > 0$ .

• Разобьем  $g$  на два сегмента:

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \begin{matrix} \} n \\ \} m_1 + m_2 - n \end{matrix}$$

Теперь после замены  $V^T y = z$  задача (22.24) переходит в эквивалентную задачу

$$\begin{bmatrix} S \\ \rho I_n \\ 0 \end{bmatrix} z \cong \begin{bmatrix} U^T d \\ \rho V^T g_1 \\ \rho g_2 \end{bmatrix}. \quad (22.26)$$

Так как система  $Cx = d$  совместна, то

$$U^T d = \begin{bmatrix} d'_1 \\ \dots \\ d'_{k_1} \\ 0 \\ \dots \\ 0 \end{bmatrix}. \quad (22.27)$$

Итак, задача (22.1) сведена к задаче (22.26), у которой матрица коэффициентов состоит из нулевой и двух диагональных матриц; одна из последних является скалярным кратным единичной матрицы.

Из лемм 22.4 и 22.7 следует, что при стремлении  $\rho$  к нулю решение  $\tilde{z}(\rho)$  задачи (22.26) сходится к (единственному) решению  $\hat{z}$  следующей задачи:

$$\text{минимизировать} \quad \|z - V^T g_1\|^2 + \|g_2\|^2 \quad (22.28)$$

при условии  $Sz = U^T d$ .

С другой стороны, посредством подстановок

$$Rx = y, \quad (22.29)$$

$$V^T y = z, \quad (22.30)$$

где  $R$  и  $V$  определены соответственно в (22.22) и (22.25), задача НКУ (20.1) преобразуется в задачу (22.28).

Поэтому, используя (22.29) и (22.30), находим, что вектор

$$\tilde{x}(\rho) = R^{-1} V \tilde{z}(\rho) \quad (22.31)$$

есть (единственное) решение задачи (22.1) при  $\rho = \epsilon(1 - \epsilon^2)^{-1/2}$ , а вектор

$$\hat{x} = R^{-1} V \hat{z} \quad (22.32)$$

есть (единственное) решение задачи НКУ (20.1).

Чтобы записать разность между  $\tilde{x}(\rho)$  и  $\hat{x}$ , заметим, что  $\tilde{z}(\rho)$  и  $\hat{z}$  связаны соотношением

$$\tilde{z}(\rho) - \hat{z} = \rho^2 h(\rho), \quad (22.33)$$

где  $h$  — векторнозначная функция, получающаяся из (22.10) при соответствующем изменении обозначений. Следовательно,

$$\tilde{x}(\rho) - \hat{x} = \rho^2 R^{-1} V h(\rho). \quad (22.34)$$

Чтобы применить к задаче (22.26) оценки (22.11), положим

$$d' = U^T d, \quad (22.35)$$

$$g' = V^T g_1. \quad (22.36)$$

Тогда из (22.11) и (22.34) выводим

$$\|\tilde{x}(\rho) - \hat{x}\| \leq \rho^2 \|R^{-1}\| s_{k_1}^{-2} \left[ \sum_{i=1}^{k_1} \left( g'_i - \frac{d'_i}{s_i} \right)^2 \right]^{1/2},$$

где  $g'_i$  и  $d'_i$  —  $i$ -е компоненты векторов (22.35) и (22.36) соответственно.



Если  $\hat{x} \neq 0$ , то для того, чтобы вектор  $\tilde{x}(\rho)$  представлял  $\hat{x}$  с относительной точностью  $\eta$  (т.е.  $\|\tilde{x}(\rho) - \hat{x}\| \leq \eta \|\hat{x}\|$ ), нужно взять положительное  $\rho$ , не превосходящее  $\rho_0$ , где

$$\rho_0^2 = \eta \|\hat{x}\| s_{k_1}^2 \left( \|R^{-1}\| \left[ \sum_{i=1}^{k_1} \left( g'_i - \frac{d'_i}{s_i} \right)^2 \right]^{1/2} \right)^{-1}.$$

Полагая

$$\epsilon_0 = \frac{\rho_0}{(1 + \rho_0^2)^{1/2}}, \quad (22.37)$$

видим, что решение  $\tilde{x}(\epsilon)$  задачи (22.1) удовлетворяет оценке

$$\|\tilde{x}(\epsilon) - \hat{x}\| \leq \eta \|\hat{x}\| \quad (22.38)$$

для  $\epsilon$  в интервале  $(0, \epsilon_0]$  \*).

Следует подчеркнуть, что трюк, состоящий в рассмотрении различных расширенных задач (22.19), (22.21), (22.24), (22.26) и (22.28) и соответствующих координатных замен, имеет целью лишь доказательство того, что при  $\epsilon \rightarrow 0$  решение задачи (22.1) сходится к решению задачи НКУ (20.1). Не нужно терять из виду, что в любой реальной вычислительной процедуре, основанной на этих идеях, можно попросту прямо решать задачу (22.1), например, с помощью преобразований Хаусхолдера.

Хотя значение  $\epsilon_0$  из формулы (22.37) дает верхнюю оценку для  $\epsilon$ , обеспечивающих полную относительную точность  $\eta$  результата, оно включает в себя величины, которые обычно при решении задачи (22.1) не вычисляются. Заметим, что ни математический анализ задачи, ни численная устойчивость метода Хаусхолдера [146] не накладывают ограничений снизу на допустимые значения  $|\epsilon|$ . Единственное практическое ограничение устанавливается значением  $L$  (см. гл. 15), характеризующим машинный нуль арифметики с плавающей запятой для данной машины.

В качестве примера рассмотрим машину, для которой  $\eta = 10^{-8}$  и  $L = 10^{-38}$ . Предположим, что все ненулевые коэффициенты матриц  $C, E$  и векторов  $d, f$  имеют приблизительно порядок 1.

С точки зрения практических запросов, скорей всего, достаточно, чтобы было  $\epsilon < \eta^{1/2} = 10^{-4}$ ; кроме того, должно быть  $\epsilon > L = 10^{-38}$ . В этом широком диапазоне можно взять, например,  $\epsilon = 10^{-12}$ .

Провилистрируем этот весовой подход к решению задачи НКУ на примере задачи (20.25) — (20.28). Сформулируем ее сейчас как взвешенную задачу наименьших квадратов

$$\begin{bmatrix} 0,4087 & 0,1593 \\ 0,4302\epsilon & 0,3516\epsilon \\ 0,6246\epsilon & 0,3384\epsilon \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \cong \begin{bmatrix} 0,1376 \\ 0,6593\epsilon \\ 0,9666\epsilon \end{bmatrix}. \quad (22.39)$$

Задача (22.39) решалась на машине UNIVAC 1108 посредством подпрограммы HFTI для 40 значений  $\epsilon$  ( $\epsilon = 10^{-r}$ ,  $r = 1, \dots, 40$ ); при этом использовался режим смешанной точности с характеристиками  $\eta = 2^{-27} \approx 10^{-8,1}$  и  $\omega = 2^{-58} \approx 10^{-17,5}$ .

\* )  $\tilde{x}(\epsilon) \equiv \tilde{x}(\rho)$ , где  $\rho = \epsilon(1 - \epsilon^2)^{-1/2}$ . (Примеч. пер.)

Таблица 22.1

| # | $\delta(r)$          | $r$ | $\delta(r)$                                                            |
|---|----------------------|-----|------------------------------------------------------------------------|
| 1 | $3,7 \times 10^{-2}$ | 10  | $5,8 \times 10^{-8}$                                                   |
| 2 | $3,7 \times 10^{-4}$ | ... | ...                                                                    |
| 3 | $3,6 \times 10^{-6}$ | 36  | $8,7 \times 10^{-8}$                                                   |
| 4 | $6,3 \times 10^{-8}$ | 37  | $3,6 \times 10^{-9}$                                                   |
| 5 | $2,6 \times 10^{-7}$ | 38  | е — слишком мало; числа, умноженные на е, превращаются в машинные нули |
| 6 | $9,1 \times 10^{-8}$ | 39  |                                                                        |
| 7 | $9,1 \times 10^{-8}$ | 40  |                                                                        |
| 8 | $1,2 \times 10^{-7}$ |     |                                                                        |
| 9 | $4,5 \times 10^{-8}$ |     |                                                                        |

"Подлинное" решение  $x$  было вычислено методом гл. 20; во всех вычислениях использовалась арифметика с удвоенной точностью ( $10^{-18}$ ). С округлением до 12 разрядов этот вектор имеет вид

$$\begin{bmatrix} -1,17749898217 \\ 3,88476983058 \end{bmatrix}.$$

Относительная точность ~~каждого приближенного~~ решения  $x^{(r)}$  вычислялась по формуле

$$\delta(r) = \frac{\|x^{(r)} - \hat{x}\|}{\|\hat{x}\|}.$$

Некоторые значения  $\delta(r)$  приведены в табл. 22.1.

Из этой таблицы видно, что любое значение  $\epsilon$  в интервале от  $10^{-4}$  до  $10^{-37}$  дает приемлемое для режима обыкновенной точности ( $\eta \approx 10^{-8,1}$ ) согласие с "подлинным" решением  $\hat{x}$ .

У п р а ж н е н и е

22.40. Пусть в (22.1) матрица  $C$  имеет размеры  $m_1 \times n$ , матрица  $E$  — размеры  $m_2 \times n$  и  $m_2 = m_1 + m_2$ . Обозначим через  $H_e$  первую матрицу Хаусхолдера, которая была бы построена при приведении к треугольному виду матрицы  $\begin{bmatrix} C \\ eE \end{bmatrix}$ . Положим

$$J_e = \begin{bmatrix} I_{m_1} & 0 \\ 0 & eI_{m_2} \end{bmatrix}.$$

Показать, что существует  $\lim_{\epsilon \rightarrow 0} J_e^{-1} H_e J_e = \tilde{H}$ . Вывести формулы, позволяющие

вычислить векторы  $u_1$  и  $u_2$  с размерностями  $m_1$  и  $m_2$  соответственно и скаляр  $b$ , (как функции первой строки матрицы  $[C^T; E^T]$ ) такие, что

$$\tilde{H} = I_m + b^{-1} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} [u_1^T \ 0].$$

Доказать, что в произведении  $\tilde{H} \begin{bmatrix} C \\ E \end{bmatrix}$  поддиагональные элементы первого столбца равны нулю. Доказать также, что соответствующую последовательность из  $m_1$  матриц типа  $\tilde{H}$  можно использовать для аннулирования всех поддиагональных элементов в первых  $m_1$  столбцах матрицы  $\begin{bmatrix} C \\ E \end{bmatrix}$ . Показать, что эта последовательность операций дает тот же результат, что и формулы (21.11) — (21.14).

# ЛИНЕЙНЫЕ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ С ЛИНЕЙНЫМИ ОГРАНИЧЕНИЯМИ-НЕРАВЕНСТВАМИ

## § 1. Введение

Существует много приложений теории наименьших квадратов в математике, физике, статистике, математическом программировании, экономике, теории управления, общественных науках и других областях, где обычную задачу НК приходится переформулировать, включив в нее некоторые ограничения в форме неравенств. В этих ограничениях заключена дополнительная информация о задаче.

Мы будем рассматривать только случай линейных неравенств. Для этой задачи в литературе предложено большое число методов. Особо отметим работы [67, 76, 174], в которых серьезное внимание уделено численной устойчивости.

Рассмотрение задач наименьших квадратов с линейными ограничениями-неравенствами позволит нам, в частности, допускать такие ограничения на решение, как неотрицательность, или независимые верхние и нижние границы для отдельных переменных, или оценки для суммы всех переменных, или (в случае сглаживания) требование, чтобы кривая была монотонной или выпуклой.

Пусть  $E, f, G, h$  — соответственно  $m_2 \times n$ -матрица,  $m_2$ -вектор,  $m \times n$ -матрица,  $m$ -вектор. Тогда задачу наименьших квадратов с линейными ограничениями-неравенствами можно сформулировать следующим образом.

**Задача НКН 23.1.** Минимизировать  $\|Ex - f\|$  при условии  $Gx \geq h$ . Будут подробно рассмотрены и важные частные случаи задачи НКН.

**Задача NNLS\*)** (наименьшие квадраты с условием неотрицательности). Минимизировать  $\|Ex - f\|$  при условии  $x \geq 0$ .

**Задача LDP\*\*) (вычисление наименьшего расстояния).** Минимизировать  $\|x\|$  при условии  $Gx \geq h$ .

Условия, характеризующие решение задачи НКН, дает теорема Куна-Таккера. Эта теорема формулируется и обсуждается в § 2 данной главы.

В § 3 исследуется задача NNLS. Представлен алгоритм ее решения, называемый NNLS. Он очень важен для последующих алгоритмов главы.

В § 4 показано, что наличие алгоритма для задачи NNLS позволяет предложить элегантное и простое решение задачи LDP.

Задача установления, совместна или нет заданная система линейных неравенств  $Cx \geq h$ , и отыскания в случае совместности какого-либо допустимого вектора возникает в различных контекстах. Разумеется, к ней можно применить алгоритм LDP. Для некоторых задач, связанных с совместностью и допустимостью, этот метод может оказаться особенно полезным ввиду отсутствия в нем ограничений на ранг и соотношение строчной и столбцовой размерностей  $G$ .

\*) NNLS — Nonnegative Least Squares. (Примеч. пер.)

\*\*) LDP — Least Distance Programming. (Примеч. пер.)

В § 5 общая задача НКН с полным столбцовым рангом преобразуется в задачу LDP. Задача НКН с дополнительными ограничениями-уравнениями рассматривается в § 6.

Наконец, в качестве приложения методов, развитых для задач с ограничениями, в § 7 разобран численный пример — сглаживание при наличии ограничений-неравенств.

## § 2. Характеризация решения

Следующая теорема характеризует вектор, решающий задачу НКН.

**Теорема 23.4.** (условия Куна-Таккера для задачи НКН). Вектор  $\hat{x}$  размерности  $n$  тогда и только тогда будет решением задачи НКН (23.1), когда найдутся  $m$ -вектор  $\hat{y}$  и такое разбиение множества  $\{1, 2, \dots, m\}$  на непересекающиеся подмножества  $\mathcal{E}$  и  $\mathcal{S}$ , что

$$G^T \hat{y} = E^T (E \hat{x} - f), \quad (23.5)$$

$$\hat{y}_i = 0, \quad i \in \mathcal{E}, \quad \hat{y}_i > 0, \quad i \in \mathcal{S}, \quad (23.6)$$

$$\hat{y}_i \geq 0, \quad i \in \mathcal{E}, \quad \hat{y}_i = 0, \quad i \in \mathcal{S}, \quad (23.7)$$

где

$$\hat{f} = G \hat{x} - h. \quad (23.8)$$

Эта теорема поддается следующей интерпретации. Пусть  $g_i^T$  —  $i$ -я строка матрицы  $G$ . Ограничение  $g_i^T x \geq h_i$  определяет допустимое полупространство  $\{x: g_i^T x \geq h_i\}$ . Вектор  $g_i$  ортогонален (нормален) к граничной гиперплоскости этого полупространства; он направлен в сторону допустимого полупространства. Точка  $\hat{x}$  является внутренней для полупространств, индексированных множеством  $\mathcal{S}$  ( $\mathcal{S}$  от слова "slack" — нежесткий), и лежит на границе полупространств, индексированных множеством  $\mathcal{E}$  ( $\mathcal{E}$  от слова "equality" — равенство).

Вектор  $p = E^T (E \hat{x} - f)$  является градиентом для функционала  $\varphi(x) = \|Ex - f\|^2/2$  в точке  $x = \hat{x}$ . Поскольку  $\hat{y}_i = 0$  для  $i \notin \mathcal{E}$ , то условие (23.5) можно переписать в виде

$$\sum_{i \in \mathcal{E}} \hat{y}_i (-g_i) = -p. \quad (23.9)$$

Смысл этого равенства такой: антиградиент  $\varphi$  в точке  $\hat{x}$  можно представить неотрицательной ( $\hat{y}_i \geq 0$ ) линейной комбинацией внешних нормалей  $(-g_i)$  к тем граничным полуплоскостям, которые содержат  $\hat{x}$ ,  $i \in \mathcal{E}$ . Геометрически это означает, что антиградиент  $(-p)$  принадлежит выпуклому конусу с вершиной в точке  $\hat{x}$ , порождаемому внешними нормальями  $(-g_i)$ .

Всякое возмущение  $u$  вектора  $\hat{x}$ , для которого вектор  $\hat{x} + u$  остается допустимым, должно удовлетворять условиям  $u^T g_i \geq 0$  для всех  $i \in \mathcal{E}$ . Умножая обе части (23.9) на такой вектор  $u$  и используя то, что  $\hat{y}_i \geq 0$ , заключаем, что  $u$  удовлетворяет также и неравенству  $u^T p \geq 0$ . Из тождества  $\varphi(\hat{x} + u) = \varphi(\hat{x}) + u^T p + \|Eu\|^2/2$  следует, что никакое допустимое возмущение  $\hat{x}$  не может уменьшить значение  $\varphi$ .

Вектор  $y$  (или противоположный вектор  $-y$ ), участвующий в формулировке теоремы Куна–Таккера, иногда называют *двойственным* вектором задачи. Обсуждение этой теоремы и ее доказательство можно найти в литературе по условной оптимизации (см., например, книгу [10]).

### § 3. Задача NNLS

Задача NNLS определена в (23.2). Мы опишем сейчас алгоритм NNLS для ее решения. Будет доказана сходимость этого алгоритма в конечное число шагов.

Поначалу заданы  $m_2 \times n$ -матрица  $E$ , целые числа  $m_2$  и  $n$  и  $m_2$ -вектор  $f$ . Векторы  $w$  и  $z$  длины  $n$  являются рабочими. В ходе выполнения алгоритма определяются индексные множества  $\mathcal{P}$  и  $\mathcal{L}$ . Переменные, индексированные множеством  $\mathcal{L}$ , имеют значение нуль. Переменные, индексированные множеством  $\mathcal{P}$ , могут принимать ненулевые значения. Если значение такой переменной неположительно, то алгоритм либо изменит его на положительное, либо, установив нулевое значение, переместит индекс переменной из множества  $\mathcal{P}$  в множество  $\mathcal{L}$ .

По окончании работы алгоритма в массиве  $x$  будет решение, а в массиве  $w$  – двойственный вектор.

Алгоритм 23.10. NNLS ( $E, m_2, n, f, x, w, z, \mathcal{P}, \mathcal{L}$ ):

1. Положить  $\mathcal{P} := \emptyset$ ,  $\mathcal{L} := \{1, 2, \dots, n\}$ ,  $x := 0$ .
2. Вычислить  $n$ -вектор  $w := E^T(f - Ex)$ .
3. Если множество  $\mathcal{L}$  пусто или  $w_j \leq 0$  для всех  $j \in \mathcal{L}$ , перейти к шагу 12.
4. Найти индекс  $i \in \mathcal{L}$  такой, что  $w_i = \max\{w_j : j \in \mathcal{L}\}$ .
5. Переместить индекс  $i$  из множества  $\mathcal{L}$  в множество  $\mathcal{P}$ .
6. Пусть  $E_{\mathcal{P}} = m_2 \times n$ -матрица, определяемая таким образом:

$$\text{столбец } j \text{ матрицы } E_{\mathcal{P}} := \begin{cases} \text{столбец } j \text{ матрицы } E, & \text{если } j \in \mathcal{P}, \\ 0, & \text{если } j \in \mathcal{L}. \end{cases}$$

Вычислить  $n$ -вектор  $z$  как решение задачи наименьших квадратов  $E_{\mathcal{P}} z \cong f$ . Заметим, что эта задача определяет только компоненты  $z_i, i \in \mathcal{P}$ . Положить  $z_i := 0$  для  $i \in \mathcal{L}$ .

7. Если  $z_i > 0$  для всех  $i \in \mathcal{P}$ , положить  $x := z$  и перейти к шагу 2.
8. Найти индекс  $q \in \mathcal{P}$  такой, что  $x_q/(x_q - z_q) = \min\{x_j/(x_j - z_j) : z_j < 0, j \in \mathcal{P}\}$ .
9. Положить  $\alpha := x_q/(x_q - z_q)$ .
10. Положить  $x := x + \alpha(z - x)$ .
11. Переместить из множества  $\mathcal{P}$  в множество  $\mathcal{L}$  все индексы  $j \in \mathcal{P}$ , для которых  $x_j = 0$ . Перейти к шагу 6.
12. Комментарий. Вычисления закончены.

Полученный алгоритмом вектор  $x$  удовлетворяет соотношениям

$$x_i > 0, \quad i \in \mathcal{P}, \quad (23.11)$$

$$x_i = 0, \quad i \in \mathcal{L}, \quad (23.12)$$

и является решением задачи наименьших квадратов

$$E_{\mathcal{P}} x \cong f. \quad (23.13)$$

Двойственный вектор  $w$  удовлетворяет соотношениям

$$w_j = 0, \quad j \in \mathcal{P}, \quad (23.14)$$

$$w_j \leq 0, \quad j \in \mathcal{L}, \quad (23.15)$$

$$w = E^T(f - Ex). \quad (23.16)$$

Соотношения (23.11), (23.12), (23.14) – (23.16) – это условия Куна–Таккера (см. теорему 23.4), характеризующие решение  $x$  задачи NNLS. Соотношение (23.13) является следствием (23.12), (23.14) и (23.16).

Обсуждению сходимости алгоритма NNLS предпослём следующую лемму.

Лемма 23.17. Пусть  $A$  –  $m \times n$ -матрица ранга  $n$ , а  $b$  –  $m$ -вектор, для которого

$$A^T b = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \omega \end{bmatrix} \begin{matrix} n-1 \\ \\ 1 \end{matrix} \quad (23.18)$$

$$\omega > 0. \quad (23.19)$$

Если  $\hat{x}$  – решение задачи  $Ax \cong b$ , то его  $n$ -я компонента

$$\hat{x}_n > 0. \quad (23.20)$$

Доказательство. Пусть  $Q$  – ортогональная  $m \times m$ -матрица, аннулирующая поддиагональные элементы в первых  $n-1$  столбцах  $A$ , т.е.

$$Q \begin{bmatrix} A & b \end{bmatrix} = \begin{bmatrix} R & s & u \\ 0 & t & v \end{bmatrix} \begin{matrix} n-1 \\ m-n+1 \end{matrix}, \quad (23.21)$$

$\begin{matrix} \overline{n} & \overline{1} & \overline{n-1} & \overline{1} & \overline{1} \end{matrix}$

где  $R$  – невырожденная верхняя треугольная матрица.

Так как  $Q$  – ортогональная матрица, то из условия (23.18) следует

$$R^T u = 0, \quad (23.22)$$

$$s^T u + t^T v = \omega > 0. \quad (23.23)$$

Поскольку  $R$  не вырождена, равенство (23.22) означает, что  $u = 0$ . Поэтому (23.23) сводится к

$$t^T v = \omega > 0. \quad (23.24)$$

Из (23.21) вытекает, что  $n$ -я компонента  $\hat{x}_n$  вектора  $\hat{x}$  сама является псевдорешением редуцированной задачи

$$t x_n \cong v. \quad (23.25)$$

Псевдообратной матрицей для столбца  $t$  является матрица  $t^T/(t^T t)$ . Поэтому решение задачи (23.25) можно записать в данном виде:

$$\hat{x}_n = \frac{t^T v}{t^T t} = \frac{\omega}{t^T t} > 0. \quad (23.26)$$

Лемма 23.17 доказана.

Можно считать, что алгоритм NNLS состоит из основного цикла (цикла  $A$ ) и внутреннего цикла (цикла  $B$ ). Цикл  $B$  составляют шаги 6–11; единственный вход в цикл  $B$  — через шаг 6, а единственный выход — через шаг 7.

Цикл  $A$  состоит из шагов 2–5 и цикла  $B$ . Он начинается шагом 2; выход из цикла — через шаг 3.

На шаге 2 множество  $\mathcal{P}$  определяет положительные компоненты текущего вектора  $x$ . Компоненты  $x$ , индексированные множеством  $\mathcal{L}$ , в этот момент равны нулю.

Выбранный на шаге 4 индекс  $i$  указывает компоненту, не представленную пока в множестве  $\mathcal{P}$ , которая в соответствии с леммой 23.17 будет положительна, если ее ввести в решение. На шаге 6 как раз и происходит ввод этой компоненты в пробное решение  $z$ . Если все прочие компоненты  $z$ , индексированные множеством  $\mathcal{P}$ , остаются положительными, то на шаге 7 выполняется присваивание  $x := z$  и возврат к началу цикла  $A$ . В рассматриваемом случае множество  $\mathcal{P}$  расширяется, а  $\mathcal{L}$  уменьшается за счет переноса индекса  $i$ .

Во многих задачах попросту происходит повторение этой последовательности событий с добавлением очередной положительной компоненты при каждом выполнении цикла  $A$ ; в конечном счете произойдет выход через шаг 3.

Если, однако, некоторая компонента, индекс которой находится в множестве  $\mathcal{P}$ , стала неположительной в векторе  $z$  шага 6, то шаг 7 вынуждает алгоритм остаться в цикле  $B$  и перейти от  $x$  к  $x + \alpha(z - x)$ ,  $0 < \alpha \leq 1$ . Множитель  $\alpha$  выбирается как можно большим; нужно лишь сохранить неотрицательность нового вектора  $x$ . Цикл  $B$  повторяется, пока в конце концов не произойдет выход через шаг 7.

Конечность цикла  $B$  можно установить, заметив, что все операции внутри этого цикла определены корректно, что при каждом выполнении шага 11 по крайней мере один индекс, а именно индекс, называемый здесь  $q$ , удаляется из множества  $\mathcal{P}$  и что  $z$ , всегда положительно (это следует из леммы 23.17). Если  $\pi$  — число индексов в множестве  $\mathcal{P}$  при входе в цикл  $B$ , то выход через шаг 7 должен произойти не позднее, чем через  $\pi - 1$  повторений цикла. На практике выход из цикла  $B$  обычно происходит при первом же достижении шага 7; при этом шаги 8–11 не выполняются вовсе.

Конечность цикла  $A$  можно установить, заметив, что норма невязки

$$\rho(x) = \|f - Ex\|$$

строго уменьшается при каждом очередном достижении шага 2. Тем самым вектор  $x$  и ассоциированное с ним множество  $\mathcal{P} = \{i: x_i > 0\}$  на шаге 2 каждый раз иные. Так как  $\mathcal{P}$  есть подмножество в  $\{1, \dots, n\}$  и таких подмножеств конечное число, то цикл  $A$  должен закончиться после конечного числа итераций. На ряде малых тестовых задач было замечено, что обычно цикл  $A$  требует около  $n/2$  итераций.

Модификация  $QR$ -разложения матрицы  $E$ . Задача наименьших квадратов, решаемая на шаге 6, отличается от задачи, решавшейся при предыдущем выполнении этого шага, либо тем, что на шаге 5 в задачу был включен дополнительный столбец, либо тем, что один или несколько столбцов  $E$  были удалены на шаге 11. Если сохранено  $QR$ -разложение преды-

дущей задачи, то для вычисления нового  $QR$ -разложения можно применить приемы модификации. Три метода модификации будут описаны в гл. 24.

Учет эффектов в машинной арифметике. Когда шаг 6 выполняется сразу вслед за шагом 5, то вычисленное на шаге 6 значение компоненты  $z_i$  теоретически должно быть положительно. Если же  $z_i$ , вследствие погрешностей округлений неположительно, то может произойти деление на ноль на шаге 8 или, возможно, будет неправильно вычислено  $\alpha = 0$  на шаге 9.

Этих неприятностей можно избежать, проверяя значение  $z_i$  после шага 6 в тех случаях, когда вход в этот шаг был из шага 5. Если окажется, что  $z_i \leq 0$ , это можно интерпретировать как указание, что значение  $w_i$ , вычисленное на шаге 2 и участвовавшее в проверках шагов 3 и 4, следует скорее считать нулем, чем положительным числом. Поэтому можно положить  $w_i := 0$  и вернуться к шагу 2. Результатом будет или выход через шаг 3, или присвоение  $i$  нового значения на шаге 4.

На шаге 11 всякое  $x_i$ , вычисленное значение которого отрицательно (что может произойти только вследствие округлений), должно рассматриваться как нулевое, а его индекс нужно переместить из множества  $\mathcal{P}$  в множество  $\mathcal{L}$ .

Проверка знаков  $z_i, i \in \mathcal{P}$ , на шагах 7 и 8, по-видимому, не является критической. Возможные погрешности в классификации не должны иметь серьезных последствий.

#### § 4. ЗАДАЧА LDP

Решение задачи LDP (23.3) можно получить путем подходящей нормировки вектора невязки соответствующей задачи NNLS (23.2). На этот метод решения задачи LDP (и на его обоснование) внимание авторов обратил Аллан Клайн.

Пусть даны  $m \times n$ -матрица  $G$ , целые числа  $m$  и  $n$  и  $m$ -вектор  $h$ . Если система неравенств  $Gx \geq h$  совместна, то логической переменной  $\varphi$  в алгоритме будет присвоено значение TRUE и будет вычислен вектор  $\hat{x}$  минимальной длины, удовлетворяющий этой системе. Если же неравенства несовместны, то алгоритм полагает  $\varphi = \text{FALSE}$ , и вектору  $\hat{x}$  не приписывается никакого значения. Рабочие массивы, необходимые алгоритму, не включены в его список параметров.

Алгоритм 23.27. LDP ( $G, m, n, h, \hat{x}, \varphi$ ):

1. Определим  $(n+1) \times m$ -матрицу  $E$  и  $n+1$ -вектор  $f$  следующим образом:

$$E = \begin{bmatrix} G^T \\ h^T \end{bmatrix}, \quad f = [0, \dots, 0, 1]^T.$$

Посредством алгоритма NNLS вычислить  $m$ -вектор  $\hat{u}$ , решающий задачу NNLS:

Минимизировать  $\|Eu - f\|$  при условии  $u \geq 0$ .

2. Вычислить  $n+1$ -вектор  $r := E\hat{u} - f$ .

3. Если  $\|r\| = 0$ , положить  $\varphi := \text{FALSE}$  и перейти к шагу 6.

4. Положить  $\varphi := \text{TRUE}$ .

5. Для  $j = 1, \dots, n$  вычислить  $\hat{x}_j := -r_j/r_{n+1}$ .

6. Вычисления закончены.

Обоснование алгоритма LDP. Рассмотрим прежде всего задачу NNLS, решаемую на шаге 1 алгоритма LDP. Для целевой функции  $\|Eu - f\|^2/2$  значением градиента в точке  $\hat{u}$  будет вектор

$$p = E^T r. \quad (23.28)$$

Согласно условиям Куна-Таккера (теорема 23.4), для этой задачи NNLS существуют непересекающиеся индексные множества  $\mathcal{E}$  и  $\mathcal{S}$  такие, что

$$\mathcal{E} \cup \mathcal{S} = \{1, 2, \dots, m\}, \quad (23.29)$$

$$\hat{u}_i = 0, \quad i \in \mathcal{E}, \quad \hat{u}_i > 0, \quad i \in \mathcal{S}, \quad (23.30)$$

$$p_i \geq 0, \quad i \in \mathcal{E}, \quad p_i = 0, \quad i \in \mathcal{S}. \quad (23.31)$$

Используя соотношения (23.28)–(23.31), получаем

$$\|r\|^2 = r^T r = r^T [E\hat{u} - f] = p^T \hat{u} - r_{n+1} = -r_{n+1}. \quad (23.32)$$

Рассмотрим случай, когда на шаге 3  $\|r\| > 0$ . Согласно (23.32), это означает, что  $r_{n+1} < 0$ , поэтому деление на  $r_{n+1}$  на шаге 5 возможно. Используя (23.31), (23.32) и соотношения шагов 2 и 5, устанавливаем допустимость вектора  $\hat{x}$ :

$$0 \leq p = E^T r = [G : h] \cdot \begin{bmatrix} \hat{x} \\ -1 \end{bmatrix} \cdot (-r_{n+1}) = (G\hat{x} - h) \|r\|^2. \quad (23.33)$$

Следовательно,

$$G\hat{x} \geq h. \quad (23.34)$$

Из (23.31) и (23.33) вытекает, что строки системы (23.34), индексированные множеством  $\mathcal{S}$ , являются в действительности точными равенствами. Градиент для целевой функции  $\|x\|^2/2$  задачи LDP – это попросту вектор  $x$ . Условия Куна-Таккера, характеризующие вектор  $\hat{x}$ , который минимизирует  $\|x\|^2/2$  при условии  $Gx \geq h$ , требуют, чтобы градиент  $\hat{x}$  был представим неотрицательной линейной комбинацией строк  $G$ , ассоциированных с равенствами в системе (23.34), т.е. строк  $G$ , индексированных множеством  $\mathcal{S}$ .

Согласно описанию шагов 2 и 5, используя (23.32), имеем

$$\hat{x} = \begin{bmatrix} r_1 \\ \dots \\ r_n \end{bmatrix} (-r_{n+1})^{-1} = G^T \hat{u} (-r_{n+1})^{-1} = G^T \hat{u} \|r\|^{-2}.$$

Условия (23.30) для знаков компонент  $\hat{u}_i$  показывают, что  $\hat{x}$  есть решение задачи LDP.

Очевидно, что это решение будет единственным. Действительно, если  $\tilde{x}$  – другое решение, то  $\|\tilde{x}\| = \|\hat{x}\|$  и вектор  $\bar{x} = (\tilde{x} + \hat{x})/2$  – допустимый вектор, длина которого строго меньше длины  $\hat{x}$ . Это противоречит тому, что  $\hat{x}$  – допустимый вектор с минимальной длиной.

Теперь рассмотрим случай, когда на шаге 3  $\|r\| = 0$ . Нужно показать, что система неравенств  $Gx \geq h$  несовместна. Предположим противное.

т.е. что существует вектор  $\tilde{x}$ , для которого  $G\tilde{x} \geq h$ . Положим

$$q = G\tilde{x} - h = [G : h] \begin{bmatrix} \tilde{x} \\ -1 \end{bmatrix} \geq 0.$$

Тогда

$$0 = [\tilde{x}^T : -1] r = [\tilde{x}^T : -1] \left( \begin{bmatrix} G^T \\ h^T \end{bmatrix} \hat{u} - f \right) = q^T \hat{u} + 1.$$

Это последнее выражение не может быть равно нулю, так как  $q \geq 0$  и  $\hat{u} \geq 0$ . Из полученного противоречия заключаем, что условие  $\|r\| = 0$  влечет за собой несовместность системы  $Gx \geq h$ . Это завершает математическое обоснование алгоритма LDP.

## § 5. Преобразование задачи НКН в задачу LDP

Рассмотрим задачу НКН (23.1) с  $m_2 \times n$ -матрицей  $E$  ранга  $n$ . Различными способами (см. гл. 2–4) можно получить ортогональное разложение матрицы  $E$

$$E = Q \begin{bmatrix} R \\ 0 \end{bmatrix} K^T \equiv \begin{bmatrix} Q_1 : Q_2 \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} K^T. \quad (23.35)$$

Здесь  $Q$  – ортогональная  $m_2 \times m_2$ -матрица,  $K$  – ортогональная  $n \times n$ -матрица и  $R$  – невырожденная  $n \times n$ -матрица. Кроме того, матрицу  $R$  можно выбрать треугольной или диагональной.

Произведем ортогональную замену переменных

$$x = Ky. \quad (23.36)$$

Целевую функцию, минимизируемую в задаче НКН, можно тогда записать в виде

$$\varphi(x) = \|f - Ex\|^2 = \left\| \begin{bmatrix} Q_1^T f \\ Q_2^T f \end{bmatrix} - \begin{bmatrix} Ry \\ 0 \end{bmatrix} \right\|^2 = \|\tilde{f}_1 - Ry\|^2 + \|\tilde{f}_2\|^2, \quad (23.37)$$

где

$$\tilde{f}_i = Q_i^T f, \quad i = 1, 2. \quad (23.38)$$

После еще одной замены переменных

$$z = Ry - \tilde{f}_1 \quad (23.39)$$

мы можем написать

$$\varphi(x) = \|z\|^2 + \|\tilde{f}_2\|^2. \quad (23.40)$$

Таким образом, исходная задача минимизации  $\|f - Ex\|$  при условии  $Gx \geq h$  эквивалентна, с точностью до аддитивной константы  $\|\tilde{f}_2\|^2$  в целе-

вой функции, следующей задаче LDP:

Минимизировать

$$\|z\| \quad (23.41)$$

при условии  $GKR^{-1}z \geq h - GKR^{-1}\tilde{f}_1$ .

Если вычислить решение  $\hat{z}$  этой задачи LDP, то решение  $\hat{x}$  исходной задачи НКН можно получить с помощью соотношений (23.39) и (23.36). Квадрат нормы невязки для исходной задачи можно вычислить по формуле (23.40).

## § 6. Задача НКН с ограничениями-уравнениями

Рассмотрим задачу НКН (23.1), к которой добавлена система ограничений-уравнений  $C_{m_1 \times n}x = d$ , где  $\text{rank } C = m_1 < n$  и  $\text{rank } [C^T : E^T] = n$ . Эти ограничения можно исключить с соответствующим сокращением числа независимых переменных. Для этой цели годятся методы гл. 20, 21.

Если использовать метод гл. 20, то выполняется ортогональная замена переменных

$$x = K \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; \quad (23.42)$$

где  $K$  — матрица, которая приводит  $C$  к треугольному виду:

$$\begin{bmatrix} C \\ E \\ G \end{bmatrix} K = \begin{bmatrix} \tilde{C}_1 & 0 \\ \tilde{E}_1 & \tilde{E}_2 \\ \tilde{G}_1 & \tilde{G}_2 \end{bmatrix}. \quad (23.43)$$

После этого из нижней треугольной системы  $\tilde{C}_1 y_1 = d$  определяется  $\hat{y}_1$ , а затем  $\hat{y}_2$  как решение следующей задачи НКН:

Минимизировать

$$\|\tilde{E}_2 y_2 - (f - \tilde{E}_1 \hat{y}_1)\| \quad (23.44)$$

при условии  $\tilde{G}_2 y_2 \geq h - \tilde{G}_1 \hat{y}_1$ .

Как только из (23.44) определен вектор  $\hat{y}_2$ , по формуле (23.42) вычисляется решение  $\hat{x}$  исходной задачи.

Если используется метод гл. 21, то по формулам (21.11)–(21.14) нужно вычислить величины  $Q_1$ ,  $\tilde{C}_1$ ,  $\tilde{C}_2$ ,  $\tilde{d}$ ,  $\tilde{E}_1$ ,  $\tilde{E}_2$ ,  $\tilde{f}$ ; кроме того, из нижних треугольных систем определяют матрицу  $\tilde{G}_1$ :

$$\tilde{G}_1 \tilde{C}_1 = G_1. \quad (23.45)$$

После этого находят вектор  $\hat{x}_2$  как решение следующей задачи НКН:

Минимизировать

$$\|\tilde{E}_2 x_2 - \tilde{f}\| \quad (23.46)$$

при условии  $(G_2 - \tilde{G}_1 \tilde{C}_2) x_2 \geq h - \tilde{G}_1 \tilde{d}$ .

Наконец, решая верхнюю треугольную систему

$$\tilde{C}_1 x_1 = \tilde{d} - \tilde{C}_2 \hat{x}_2, \quad (23.47)$$

вычисляют вектор  $\hat{x}_1$ .

## § 7. Пример выравнивания при наличии ограничений

В качестве примера, который иллюстрирует ряд приемов, описанных в этой главе, рассмотрим задачу выравнивания экспериментальных данных при некоторых ограничениях на аппроксимирующую прямую.

В следующей таблице приведены исходные данные задачи:

| $t$  | $w$ | $t$  | $w$ |
|------|-----|------|-----|
| 0,25 | 0,5 | 0,50 | 0,7 |
| 0,50 | 0,6 | 0,80 | 1,2 |

Мы хотим найти прямую

$$f(t) = x_1 t + x_2, \quad (23.48)$$

которая аппроксимирует эти данные методом наименьших квадратов при следующих ограничениях:

$$f'(t) \geq 0, \quad (23.49)$$

$$f(0) \geq 0, \quad (23.50)$$

$$f(1) \leq 1. \quad (23.51)$$

Эту задачу выравнивания можно представить в форме задачи НКН:

Минимизировать

$$\|Ex - f\| \quad (23.52)$$

при условии  $Gx \geq h$ , где

$$E = \begin{bmatrix} 0,25 & 1 \\ 0,50 & 1 \\ 0,50 & 1 \\ 0,80 & 1 \end{bmatrix}, \quad f = \begin{bmatrix} 0,5 \\ 0,6 \\ 0,7 \\ 1,2 \end{bmatrix},$$

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}, \quad h = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

Чтобы (как это описано в § 5) свести задачу НКН к задаче LDP, вычислим ортогональное разложение матрицы  $E$ . Годятся и сингулярное разложение  $E$ , и QR-разложение  $E$ . Проиллюстрируем применение сингуляр-

ного разложения:

$$E = U_4 \times 4 \begin{bmatrix} S_{2 \times 2} \\ 0_{2 \times 2} \end{bmatrix} V_{2 \times 2}^T,$$

$$S = \begin{bmatrix} 2,255 & 0,0 \\ 0,0 & 0,346 \end{bmatrix}, V = \begin{bmatrix} -0,467 & 0,884 \\ -0,884 & -0,467 \end{bmatrix},$$

$$2 \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix} = U^T f = \begin{bmatrix} -1,536 \\ 0,384 \\ -0,054 \\ 0,174 \end{bmatrix}.$$

Выполним замену переменных

$$z = SV^T x - \tilde{f}_1. \quad (23.53)$$

Теперь нужно решить следующую задачу LDP:  
Минимизировать

$$\|z\| \quad (23.54)$$

при условии  $\tilde{G}z \geq \tilde{h}$ , где

$$\tilde{G} = GVS^{-1} = \begin{bmatrix} -0,207 & 2,558 \\ -0,392 & -1,351 \\ 0,599 & -1,206 \end{bmatrix}, \quad \tilde{h} = h - \tilde{G}\tilde{f}_1 = \begin{bmatrix} -1,300 \\ -0,084 \\ 0,384 \end{bmatrix}.$$

Графическая интерпретация этой задачи дана на рис. 23.1. Каждая строка расширенной матрицы  $[\tilde{G} : \tilde{h}]$  определяет граничную прямую области

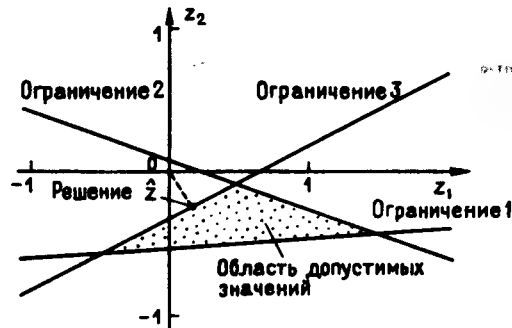


Рис. 23.1. Графическая интерпретация иллюстративной задачи LDP (23.54)

допустимых значений. Решение  $\hat{x}$  есть точка этой области, ближайшая к началу координат. Вычисления по алгоритму LDP дают

$$\hat{z} = \begin{bmatrix} 0,127 \\ -0,255 \end{bmatrix}.$$

Рис. 23.2. График выравнивающей прямой для иллюстративной задачи (23.48) – (23.51)

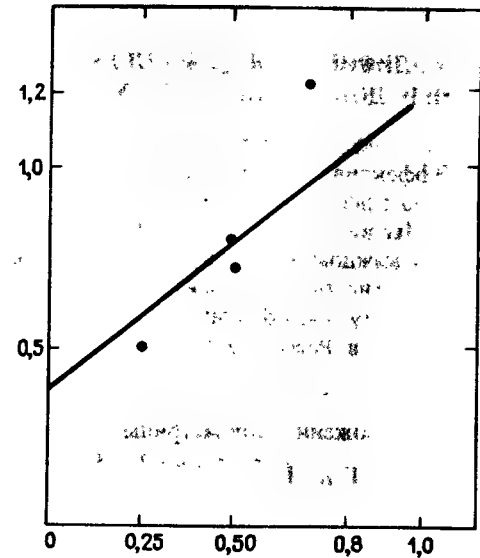
Теперь, согласно (23.53), находим

$$\hat{x} = VS^{-1}(\hat{z} + \tilde{f}_1) = \begin{bmatrix} 0,621 \\ 0,379 \end{bmatrix}.$$

Невязка для решения  $\hat{x}$  имеет вид

$$\hat{r} = f - E\hat{x} = \begin{bmatrix} -0,034 \\ -0,089 \\ 0,011 \\ 0,324 \end{bmatrix},$$

$$\|\hat{r}\| = 0,338.$$



На рис. 23.2 показаны заданные точки  $(t_i, w_i), i = 1, \dots, 4$ , и выравнивающая прямая  $f(t) = 0,621t + 0,379$ . Отметим, что третье ограничение  $f(1) \leq 1$  оказывается активным и влияет на точность аппроксимации заданных точек.

Числовые значения, указанные в разборе данного примера, были получены на машине UNIVAC 1108. Когда та же фортранная программа выполнялась на IBM 360/67, то промежуточные величины  $V, \tilde{f}_1, \tilde{G}, \hat{z}$  были вычислены с противоположными знаками. Это является следствием того обстоятельства, что знаки столбцов матрицы  $V$  из сингулярного разложения не определены однозначно. Разное число итераций, потребовавшееся при вычислении сингулярного разложения матрицы  $E$  на двух машинах с различной длиной слова, привело к разному приписыванию знаков матрицам  $U$  и  $V$ .

Упражнения

23.55. Доказать, что если задача НКУ имеет единственное решение без ограничений-неравенств, то она имеет единственное решение и при наличии таких ограничений\*).

23.56. Показать, что задача минимизации квадратичной функции  $f(x) = x^T Bx/2 + a^T x$ , где  $B$  – положительно определенная матрица, заменой переменных  $w = Fx - g$  и подходящим выбором невырожденной матрицы  $F$  и вектора  $g$  сводится к задаче минимизации  $\|w\|^2/2$ .

23.57. Если функцию  $f(x)$  из упражнения 23.56 нужно минимизировать при наличии ограничений  $Cx = d$  и  $Gx \geq h$ , то каковы будут соответствующие ограничения в эквивалентной задаче минимизации  $\|w\|^2/2$ ?

\*) Разумеется, в случае совместности с дополнительными ограничениями. (Примеч. пер.)

# МОДИФИКАЦИЯ QR-РАЗЛОЖЕНИЯ МАТРИЦЫ ПРИ ДОБАВЛЕНИИ ИЛИ УДАЛЕНИИ СТОЛБЦОВ

Эффективность алгоритма гл. 23 для задачи НКН связана с возможностью эффективно решать последовательность задач наименьших квадратов. Мы видели, что в алгоритме NNLS (см. 23.10) матрица коэффициентов на каждом шаге получается из предыдущей добавлением или удалением столбца (не являющегося линейной комбинацией остальных).

Поэтому мы обсудим такую задачу. Пусть  $n > 0$  и  $a_1, \dots, a_n$  — система  $m$ -векторов. Рассмотрим матрицу

$$A_k = [a_1 \dots a_k] \quad (24.1)$$

и QR-разложение этой матрицы

$$QA_k = \begin{bmatrix} R_k \\ 0 \end{bmatrix}, \quad (24.2)$$

где  $R_k$  — порядка  $k$  невырожденная верхняя треугольная матрица\*), а  $Q$  — ортогональная матрица. Как только  $Q$  и  $R_k$  вычислены, решение задачи

$$A_k x \cong b \quad (24.3)$$

можно найти по формуле

$$x = [R_k^{-1} : 0] Q^T b \equiv A_k^+ b, \quad (24.4)$$

требующей лишь очень небольшой дополнительной работы.

Рассмотрим задачу о вычислении QR-разложения матрицы, полученной из  $A_k$  удалением или приписыванием столбца, и постараемся извлечь выгоду из имеющегося QR-разложения матрицы  $A_k$ . Опишем три полезных метода модификации QR-разложения. Как показывает (24.4), модификация  $Q$  и  $R$ , по существу, дает способ модификации матрицы  $A_k^+$ .

Метод 1. Предположим, что  $Q$  хранится как явно вычисленная\*\*) ортогональная  $m \times m$ -матрица, а  $R$  — как верхняя треугольная  $k \times k$ -матрица.

Приписывание вектора. Пусть  $a_{k+1}$  — вектор, не являющийся линейной комбинацией столбцов  $A_k$ . Составим расширенную матрицу

$$A_{k+1} = [A_k : a_{k+1}]. \quad (24.5)$$

Вычислим произведение  $b_{k+1} = Q a_{k+1}$ , где  $Q$  — матрица из (24.2). Построим матрицу Хаусхолдера  $Q_{k+1}$  так, чтобы вектор  $r_{k+1} = Q_{k+1} b_{k+1}$  имел нулевые компоненты в позициях  $k+2, \dots, m$ \*\*\*). Тогда QR-разложение матрицы  $A_{k+1}$  дает формула

$$\tilde{Q} A_{k+1} = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix},$$

\*) Система  $a_1, \dots, a_n$  считается линейно независимой. (Примеч. пер.)

\*\*) А не факторизованная, как в методе 2. (Примеч. пер.)

\*\*\*) Сохранив неизменными компоненты с номерами  $1, \dots, k$ . (Примеч. пер.)

где

$$\tilde{Q} = Q_{k+1} Q, \quad \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix} : r_{k+1}.$$

Удаление вектора. Пусть матрица  $A_k$  из (24.1) модифицируется удалением столбца  $a_j$ , в результате чего получается матрица

$$A_{k-1} = [a_1 \dots a_{j-1} a_{j+1} \dots a_k]. \quad (24.6)$$

Если положить

$$Q A_{k-1} = [r_1 \dots r_{j-1} r_{j+1} \dots r_k], \quad (24.7)$$

то каждый вектор  $r_i$  имеет нули в позициях  $i+1, \dots, m$  (см. [76, 174]). Введем матрицу

$$\hat{Q} = G_{k-1} \dots G_j Q, \quad (24.8)$$

где  $G_i$  — вращения Гивенса, выбранные так, чтобы матрица

$$\hat{Q} A_{k-1} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} \quad (24.9)$$

была верхней треугольной. Матрица  $G_i$  оперирует со строками  $i$  и  $i+1$  и позволяет получить нуль в позиции  $(i+1, i)$  произведения  $G_i \dots G_j Q A_{k-1}$ .

Метод 1 требует  $m^2$  машинных слов для хранения  $Q$  и  $l(l+1)/2$  слов для хранения  $R$ . Здесь  $l, l \leq m$ , — наибольшее число векторов  $a_s$ , используемых одновременно.

Метод 2. Вместо явного хранения матрицы  $Q$ , как в методе 1, мы будем теперь считать, что  $Q$  задана как произведение  $k$  преобразований Хаусхолдера

$$Q = Q_k \dots Q_1. \quad (24.10)$$

Для каждой матрицы

$$Q_i = I_m + b_i^{-1} u_i u_i^T \quad (24.11)$$

хранятся лишь ненулевые элементы вектора  $u_i$ , а также используется еще одна ячейка. Матрица  $R$  по-прежнему хранится как верхняя треугольная  $k \times k$ -матрица.

Приписывание вектора. Пусть система столбцов (24.1) расширена с сохранением свойства линейной независимости до системы  $[a_1 \dots a_{k+1}]$ . Вычислим произведение  $b_{k+1} = Q_k \dots Q_1 a_{k+1}$ . Построим преобразование Хаусхолдера  $Q_{k+1}$  так, чтобы вектор  $Q_{k+1} b_{k+1} = r_{k+1}$  имел нули в позициях  $k+2, \dots, m$ \*\*\*). Тогда формула

$$Q_{k+1} (Q_k \dots Q_1) [A_k : a_{k+1}] = \begin{bmatrix} R \\ 0 \end{bmatrix} : r_{k+1} \quad (24.12)$$

дает QR-разложение расширенной матрицы.

\*) См. примечание на стр. 134 (Примеч. пер.)



Удаление вектора. Снова рассмотрим матрицу  $A_{k-1}$  из (24.6). Запишем более подробно формулу, определяющую  $QR$ -разложение матрицы  $A_k$ :

$$Q_k \dots Q_1 A_k = \begin{bmatrix} R \\ 0 \end{bmatrix} = \begin{bmatrix} \overbrace{R'_{11} \dots R'_{1j-1}}^{j-1} & \overbrace{R'_{1j} \dots R'_{1k}}^{k-j+1} \\ 0 & \overbrace{R'_{22} \dots R'_{2k}}^{k-j+1} \\ 0 & 0 \end{bmatrix} \begin{matrix} j-1 \\ k-j+1 \\ m-k \end{matrix} \quad (24.13)$$

Представим искомое  $QR$ -разложение  $A_{k-1}$  в виде

$$Q A_{k-1} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}.$$

В матрицу  $\hat{R}$  в качестве ведущей главной подматрицы порядка  $j-1$  входит та же, что и в (24.13), матрица  $R_{11}$ . Поэтому нужно найти только последние  $k-j$  столбцов  $\hat{R}$ .

Одна из возможностей состоит в том, чтобы на место подматрицы

$$\begin{bmatrix} R'_{12} \\ R'_{22} \\ 0 \end{bmatrix}$$

записать систему  $[a_{j+1} \dots a_k]$ . Далее вычисляем произведение

$$Q_{j-1} \dots Q_1 [a_{j+1} \dots a_k] = S_{m \times (k-j)}. \quad (24.14)$$

Наконец, строим новые преобразования Хаусхолдера  $Q'_i$  так, чтобы

$$Q'_{k-1} \dots Q'_j S = \begin{bmatrix} \overbrace{\hat{R}_{12} \dots \hat{R}_{1k-j}}^{k-j} \\ \overbrace{\hat{R}_{22} \dots \hat{R}_{2k-j}}^{k-j} \\ 0 \end{bmatrix} \begin{matrix} j-1 \\ k-j \\ m-k+1 \end{matrix},$$

где  $\hat{R}_{22}$  — верхняя треугольная матрица.

$QR$ -разложение матрицы  $A_{k-1}$  теперь имеет вид

$$Q'_{k-1} \dots Q'_j Q_{j-1} \dots Q_1 A_{k-1} = \begin{bmatrix} R_{11} & \hat{R}_{12} \\ 0 & \hat{R}_{22} \\ 0 & 0 \end{bmatrix}.$$

Иначе матрицу  $S$  из (24.14) можно построить по формуле

$$Q_j Q_{j+1} \dots Q_k \begin{bmatrix} R'_{12} \\ R'_{22} \\ 0 \end{bmatrix} = [\tilde{a}_j : S], \quad \tilde{a}_j = Q_{j-1} \dots Q_1 a_j, \quad (24.15)$$

где  $R'_{12}$  и  $R'_{22}$  определены в (24.13).

Метод 2 требует  $(m+1)l$  машинных слов для хранения информации, определяющей  $Q$  и  $R$ . Здесь  $l$ ,  $l \leq m$ , — наибольшее число векторов  $a_s$ , используемых одновременно. Вариант метода 2, опирающийся на формулу (24.14), требует дополнительного  $m \times n$ -массива для хранения копий векторов  $a_1, \dots, a_n$ , с тем чтобы при необходимости использовать их в вычислениях по этой формуле.

Во втором варианте метода 2, использующем соотношение (24.15), устраняется необходимость в хранении копий векторов  $a_j, a_{j+1}, \dots, a_n$ . Поэтому вся информация о  $QR$ -разложении матрицы  $A_k$  и векторы  $a_i$ ,

не включаемые в данный момент в число столбцов  $A_k$ , могут быть (с точностью до дополнительных  $3m$  ячеек) храниться единым  $m \times n$ -массивом.

Метод 3. В этом методе все содержимое массива  $W_{m \times (n+1)}$ , который поначалу хранит исходные данные задачи  $[A_{m \times n} : b_{m \times 1}]$ , модифицируется каждый раз, когда приписывается или удаляется из базиса очередной вектор. Столбцы верхней треугольной матрицы  $R_k$  из (24.2) занимают некоторые  $k$  столбцов массива  $W$ . Матрица  $Q$  из этой формулы не хранится.

Пусть  $\mathcal{P}_k = \{p_1, \dots, p_k\}$  — подмножество в  $\{1, \dots, n\}$ , определяющее столбцы  $A$ , введенные в базис. Порядок чисел  $p_i$  в множестве  $\mathcal{P}_k$  важен. Пусть  $A_k$  —  $m \times k$ -матрица, образованная столбцами  $A$ , которые индексированы множеством  $\mathcal{P}$  в порядке  $p_1, \dots, p_k$ . Пусть  $Q$  — ортогональная матрица из формулы (24.2). Тогда в массиве  $W$  находится  $m \times (n+1)$ -матрица  $Q[A : b]$ . Это значит, что  $j$ -й столбец  $R_k$  хранится в столбце  $p_j$  массива  $W$ . В рассматриваемом методе может оказаться удобным явное хранение нулевых элементов, полученных преобразованиями Хаусхолдера или Гивенса.

Приписывание вектора. Пусть  $p_{k+1}$  — индекс столбца  $A$ , приписываемого к базису. Составим новое индексное множество  $\mathcal{P}_{k+1}$ , присоединяя к  $\mathcal{P}_k$  индекс  $p_{k+1}$ . Построим преобразование Хаусхолдера  $H$ , которое в столбце  $p_{k+1}$  массива  $W$  аннулирует элементы, находящиеся ниже строки  $k+1$ . Умножим на  $H$  слева весь массив  $W$ .

Удаление вектора. Снова предположим, что  $\mathcal{P}_k$  определяет текущий базис. Пусть из базиса нужно удалить столбец  $p_j$ ,  $1 \leq j \leq k$ . Для  $i = j+1, \dots, k$  построим вращение Гивенса  $G_i$ , оперирующее со строками  $i-1$  и  $i$  и аннулирующее элемент  $W$  в позиции  $(i, p_i)$ . Умножим на  $G_i$  слева весь массив  $W$ .

Образуем новое индексное множество  $\mathcal{P}_{k-1}$ , полагая  $p_i := p_i$ ,  $i = 1, \dots, j-1$ , и  $p_i := p_{i+1}$ ,  $i = j, \dots, k-1$ . В этом методе требуется лишь  $m \times (n+1)$ -массив  $W$ , содержащий поначалу исходные данные  $[A : b]$ . Если на последующих этапах вычислений потребуются копии этих данных, то для их хранения должен быть выделен дополнительный массив.

## ГЛАВА 25

### ПРАКТИЧЕСКИЙ АНАЛИЗ ЗАДАЧ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

#### § 1. Общие соображения

В этой главе мы обсудим стратегию планирования и интерпретации при практическом решении задач метода наименьших квадратов. Мы будем рассматривать случай  $m \geq n$ , который является центральным для нашей книги.

Пользователь, от которого исходит задача наименьших квадратов (имеваемый в дальнейшем хозяином задачи), должен обзавестись начальными

\*) См. примечание на стр. 134 (Примеч. пер.)

данными и построить или выбрать математическую модель, а часто также и модель статистическую. В какой-то момент своей работы он достигает этапа, когда у него имеются матрица  $A$  и вектор  $b$ , и нужно найти вектор  $x$ , который минимизировал бы евклидову норму вектора невязки

$$r = b - Ax. \quad (25.1)$$

Хозяин задачи должен еще иметь информацию о неопределенности в задании коэффициентов матрицы  $A$  и вектора  $b$ . Зачастую у него есть и априорная информация относительно решения задачи. Она может состоять в примерном предварительном знании области разумных значений для некоторых или всех компонент решения. Другой возможный вариант предварительной информации — требование, чтобы все или часть компонент были неотрицательны.

С очень общей точки зрения имеются два главных соображения, которые нужно учитывать при численном решении задачи. Важно разделить их.

1. Вычислительная погрешность может быть снижена до уровня, когда она является пренебрежимо малой в сравнении с неопределенностью решения, вызванной неопределенностью во входных данных задачи.

2. Комбинированный эффект априорной неопределенности в  $A$  и  $b$  и обусловленности матрицы  $A$  может привести к ситуации, когда имеется ряд существенно различающихся векторов  $x$ , которые дают для нормы невязки  $r$  приемлемо малое значение. Мы обсудим приемы, позволяющие обнаружить такую ситуацию, и некоторые методы управления выбором конкретного решения из множества решений-кандидатов.

Остановимся более подробно на соображении 1. Предположим, что имеющаяся информация о неопределенности в  $A$  и  $b$  может быть выражена следующим утверждением. Известны числа  $\varphi$  и  $\psi$  такие, что всякая матрица вида  $A + E$ , где

$$\|E\| \leq \varphi, \quad (25.2)$$

и всякий вектор вида  $b + db$ , где

$$\|db\| \leq \psi, \quad (25.3)$$

приемлемы для хозяина задачи в качестве замены наличных входных данных  $A$  и  $b$ .

Для сопоставления с этими параметрами  $\varphi$  и  $\psi$  неустранимой погрешности введем, пользуясь оценками (16.3) и (16.4), следующие две функции от  $\eta$ :

$$\varphi'(\eta) = (6m - 3n + 41)n^{3/2} \|A\| \eta, \quad (25.4)$$

$$\psi'(\eta) = (6m - 3n + 40)n \|b\| \eta. \quad (25.5)$$

Мы заменили в (16.3) величину  $\|A\|_F$  ее верхней оценкой  $n^{1/2} \|A\|$ .

Игнорируя в (16.3) и (16.4) члены  $O(\eta^2)$ , мы извлекаем из теоремы 16.1 такой вывод: если  $\eta$  выбрано настолько малым, чтобы

$$\varphi'(\eta) < \varphi, \quad (25.6)$$

$$\psi'(\eta) < \psi, \quad (25.7)$$

то решение, вычисленное в арифметике с постоянной точностью  $\eta$ , явля-

ется точным для некоторой задачи, исходные данные которой отличаются от заданных  $A$  и  $b$  возмущениями, удовлетворяющими оценкам (25.2) и (25.3).

Аналогично в случае вычислений со смешанной точностью для выбора  $\eta$  и  $\omega (\leq \eta^2)$  мы можем с помощью (17.13) и (17.14) определить функции

$$\varphi''(\eta) = 30n^{3/2} \|A\| \eta, \quad (25.8)$$

$$\psi''(\eta) = 29n \|b\| \eta. \quad (25.9)$$

Возьмем  $\eta$  такое, чтобы

$$\varphi''(\eta) < \varphi, \quad (25.10)$$

$$\psi''(\eta) < \psi. \quad (25.11)$$

При этом выборе  $\eta$  и  $\omega \leq \eta^2$  мы заключаем по теореме 17.11, что вычисленное решение является точным для возмущенной задачи, причем возмущения будут меньше, чем априорная неопределенность, описываемая оценками (25.2) и (25.3).

Напомним, что множители перед произведениями  $\|A\| \eta$  и  $\|b\| \eta$  в формулах (25.4), (25.5), (25.8) и (25.9) были получены в предположении о наихудшем накоплении погрешностей округлений. Как показывает практика, замена этих множителей квадратными корнями из них дает обычно более реалистическое представление о величине возмущений, эквивалентных вычислительным погрешностям.

Перейдем теперь к соображению 2. Полезно формализовать понятие "приемлемо малого" вектора невязки. Предположим, что мы готовы принять невязки с нормами, не превосходящими некоторого числа  $\rho$ . Тогда можно определить множество приемлемых решений так:

$$X = \{x: \|(A + E)x - (b + db)\| \leq \rho, \|E\| \leq \varphi, \|db\| \leq \psi\}. \quad (25.12)$$

Важно отметить, что определение множества  $X$  опирается на три допущения  $\varphi$ ,  $\psi$  и  $\rho$ , которые следует выбирать, исходя из имеющейся информации о задаче и ее входных данных.

Формулируя (25.12), мы ставим своей целью не столько утвердить именно это определение, сколько придать некоторую степень конкретности общей идее множества приемлемых решений, а также указать величины, от которых зависит "размер" этого множества.

Чтобы оценить "размер" множества  $X$ , заметим прежде всего, что  $X$  будет не ограничено, если  $\text{rank } A < n$  или  $\varphi$  и  $\kappa$  (где  $\kappa$  обозначает число обусловленности  $A$ ) настолько велики, что  $\kappa\varphi/\|A\| > 1$ ; в этом последнем случае  $A + E$  будет вырождена для некоторой матрицы  $E$ ,  $\|E\| \leq \varphi$ .

С другой стороны, в случае  $\text{rank } A = n$  и  $\kappa\varphi/\|A\| < 1$  оценки возмущений (9.13) или (9.14) позволяют получить полезную информацию о диаметре множества  $X$ . Присутствие параметра  $\rho$  в определении  $X$  может повести к дальнейшему увеличению его размера.

Если множество  $X$  "велико" в том смысле, что содержит векторы, значительно отличающиеся друг от друга, то приходится выбирать какое-либо конкретное решение из  $X$ . При широком взгляде на дело в этот процесс выбора можно включить любые шаги, которые может предпринять хозяин задачи, с тем чтобы путем ее преобразования уменьшить множество решений (сохраняя, как правило, вложенность в  $X$ ).

Критерий, используемый для уменьшения размера множества  $X$ , зависит от конкретного приложения. Очень распространена ситуация, когда задача  $Ax \cong b$  получается в результате локальной линеаризации нелинейной задачи наименьших квадратов. Это уже отмечалось в гл. 1. В подобном случае предпочтительным обычно является выбор из  $X$  вектора  $\hat{x}$  с наименьшей нормой; такой выбор снижает вероятность выхода из области, где  $b - Ax$  — хорошее приближение к нелинейной задаче.

Что касается преобразования исходной задачи наименьших квадратов, то было предложено множество конкретных процедур, опирающихся на различные мотивировки (статистические, математические, численные, эвристические и т.д.). Большинство этих процедур заключается в выполнении одной или более из операций следующего списка (не обязательно в указанном порядке):

1. Левое умножение  $A$  и  $b$  на  $m \times m$ -матрицу  $G$ .
2. Правое умножение  $A$  на  $n \times n$ -матрицу  $H$  и соответствующая замена переменных  $x = H\tilde{x}$  или  $x = H\tilde{x} + \xi$ .
3. Приписывание дополнительных строк к  $A$  и дополнительных элементов к  $b$ .
4. Присвоение фиксированных значений (часто нулевых) некоторым компонентам решения. Это можно сделать по отношению как к исходному, так и к преобразованному набору переменных.

В § 2–5 мы рассмотрим подробно каждую из этих четырех операций.

Наконец, в § 6 будет описан сингулярный анализ. Под этим термином мы подразумеваем вычисление ряда величин, которые можно извлечь из сингулярного разложения матрицы  $A$ , а также их интерпретацию как средство, чтобы достигнуть понимания неопределенностей в задаче  $Ax \cong b$  и выбрать полезное решение. Разумеется, задача  $Ax \cong b$ , к которой применяется сингулярный анализ, может быть получена в результате предварительного использования операций, описываемых в § 2–5.

## § 2. Левое умножение $A$ и $b$ на матрицу $G$

Эта операция изменяет норму, в которой оценивается величина вектора невязки. Таким образом, задачу выбора  $x$  из условия минимума  $(b - Ax)^T(b - Ax)$  мы заменяем на задачу минимизации величины  $(Gb - GAx)^T(Gb - GAx)$ , которую можно записать также как  $(b - Ax)^T \times \times (G^T G)(b - Ax)$  или как  $(b - Ax)^T W(b - Ax)$ , где  $W = G^T G$ .

Очень часто используется специальный случай, когда  $G$  — диагональная матрица. Тогда и матрица  $W$  диагональная. В этом случае левое умножение на  $G$  можно интерпретировать как операцию масштабирования строк, в которой  $i$ -я строка расширенной матрицы  $[A : b]$  умножается на число  $g_{ii}$ .

Положим

$$r = b - Ax, \quad (25.13)$$

$$\tilde{r} = Gr = Gb - GAx. \quad (25.14)$$

Таким образом, если матрица  $G$  диагональная, то нужно минимизировать величину

$$\|\tilde{r}\|^2 = \sum_{i=1}^m g_{ii}^2 r_i^2 = \sum_{i=1}^m w_{ii} r_i^2. \quad (25.15)$$

Говоря нестрого, приписывание  $i$ -му уравнению относительно большого веса  $|g_{ii}|$  (или, что эквивалентно,  $w_{ii}$ ) имеет тенденцию уменьшать компоненту  $|r_i|$  в результирующей невязке. Поэтому если некоторые компоненты исходного вектора  $b$  известны с большей абсолютной точностью, чем другие, то может оказаться желательным введение для соответствующих строк сравнительно больших весов.

Описанную процедуру для диагональной матрицы  $G$  обычно называют взвешенными наименьшими квадратами. Чтобы дать систематический метод для приписывания весов, предположим, что с каждой компонентой  $b_i$  исходного вектора  $b$  можно связать положительное число  $\sigma_i$ , указывающее приблизительный размер неопределенности в  $b_i$ . Если имеется соответствующая статистическая информация относительно  $b$ , то обычно в качестве  $\sigma_i$  берут среднееквадратичное отклонение неопределенности в  $b_i$ . Тогда веса определяются, как правило, формулой  $g_{ii} = 1/\sigma_i$  или, что эквивалентно,  $w_{ii} = 1/\sigma_i^2$ . Заметим, что при таком масштабировании все компоненты модифицированного вектора  $\tilde{b}$ , т.е.  $\tilde{b}_i = g_{ii} b_i = b_i/\sigma_i$ , имеют неопределенности со среднееквадратичным отклонением 1.

Более общо, если имеется достаточная статистическая информация относительно неопределенности в векторе  $b$  для того, чтобы приписать числовые значения корреляции погрешностей в различных его компонентах, то эту информацию можно представить в виде положительно определенной симметричной ковариационной  $m \times m$ -матрицы  $C$  (эти статистические понятия подробно обсуждаются в книге [144]). Тогда можно вычислить разложение Холецкого матрицы  $C$  в соответствии с формулами (19.5) и (19.12)–(19.14), т.е. найти нижнюю треугольную матрицу  $F$ , для которой  $C = FF^T$ . После этого весовую матрицу  $G$  можно определить как  $G = F^{-1}$ .

Нет необходимости вычислять обратную для  $F$  в явном виде. Более экономичный метод состоит в прямом вычислении взвешенной матрицы  $[A : \tilde{b}]$  путем решения систем  $F[A : \tilde{b}] = [A : b]$  с треугольной матрицей  $F$ .

Если  $G$  получена описанным способом из априорной ковариационной матрицы погрешностей в  $b$ , то ковариационная матрица погрешностей в преобразованном векторе  $\tilde{b}$  будет единичной.

Как статистические, так и всякого рода иные соображения свидетельствуют в пользу выбора матрицы  $G$ , обеспечивающей примерное равенство областей неопределенности для всех компонент преобразованного вектора  $\tilde{b} = Gb$ . В результате евклидова норма становится разумной мерой вектора погрешности  $db$  из (25.3).

## § 3. Правое умножение $A$ на матрицу $H$ и замена переменных $x = H\tilde{x} + \xi$

Здесь задача

$$Ax \cong b \quad (25.16)$$

заменяется задачей

$$\tilde{A}\tilde{x} \cong \tilde{b}, \quad (25.17)$$

где

$$\tilde{A} = AH, \quad (25.18)$$

$$\tilde{b} = b - A\xi, \quad (25.19)$$

$$x = H\tilde{x} + \xi. \quad (25.20)$$

Матрица  $H$  имеет размеры  $n \times l$ , причем  $l \leq n$ . Вектор  $\tilde{x}$  —  $l$ -мерный, а  $\tilde{A}$  —  $m \times l$ -матрица. Если  $H$  — диагональная  $n \times n$ -матрица, то такое преобразование можно интерпретировать как применение к  $A$  операции масштабирования столбцов.

Для невырожденной  $n \times n$ -матрицы  $H$  преобразование не меняет задачу математически. Таким образом, множество векторов вида  $x = H\tilde{x} + \xi$ , где  $\tilde{x}$  минимизирует  $\|b - \tilde{A}\tilde{x}\|$ , совпадает с множеством векторов  $x$ , минимизирующих  $\|b - Ax\|$ .

Если, однако,  $H$  — неортогональная матрица, то число обусловленности  $\tilde{A}$ , вообще говоря, отличается от числа обусловленности  $A$ . Поэтому если используется алгоритм, в котором нужно определять значение псевдоранга  $A$  (например, алгоритм HFTI (см. 14.9) или алгоритм сингулярного разложения (18.36)–(18.45)), то для матрицы  $\tilde{A}$  это значение может быть выбрано иным.

Кроме того, если для псевдоранга установлено значение  $k$ , меньшее  $n$ , а затем вычисляется решение с минимальной длиной для задачи ранга  $k$ , то использование неортогональной матрицы  $H$  изменяет норму, посредством которой измеряется "величина" решения. В общем случае это приводит к тому, что в качестве решения с "минимальной длиной" выбирается другой вектор. Если в задаче (25.16) решение с минимальной длиной — это решение, минимизирующее  $\|x\|$ , то в преобразованной задаче (25.17) решением с минимальной длиной будет решение, минимизирующее  $\|\tilde{x}\|$ . Это равносильно минимизации  $\|H^{-1}(x - \xi)\|$  вместо минимизации  $\|x\|$ .

По поводу критерия для выбора  $H$  заметим прежде всего, что использование спектральной нормы для матрицы возмущения  $E$  в (25.2) будет реалистичным только тогда, когда абсолютные неопределенности всех элементов  $A$  имеют примерно одинаковую величину. Поэтому если имеются индивидуальные оценки неопределенностей для элементов  $A$ , то матрицу  $H$  можно выбирать как матрицу, масштабирующую столбцы  $A$ , с тем чтобы уравновесить величины неопределенностей в них.

Сходный критерий можно получить на основе априорной информации о решении  $x$ . Предположим, известно, что решение  $x$  должно быть близко к заданному вектору  $\xi$  (априорно ожидаемому значению  $x$ ). Пусть далее для каждого  $i$  имеется априорная оценка  $\sigma_i$  неопределенности в  $\xi_i$  как оценки  $x_i$ . Тогда в качестве  $H$  можно взять диагональную  $n \times n$ -матрицу с диагональными элементами

$$h_{ii} = \sigma_i. \quad (25.21)$$

В таком случае преобразованные переменные

$$\tilde{x}_i = \frac{x_i - \xi_i}{\sigma_i} \quad (25.22)$$

имеют единичную априорную неопределенность и нулевые априорно ожидаемые значения.

Более общо, если имеется достаточная априорная статистическая информация для того, чтобы построить положительно определенную априорную ковариационную  $n \times n$ -матрицу  $C$ , описывающую неопределенность в  $\xi$ , то  $H$  можно вычислить как верхний треугольный множитель Холецкого этой матрицы (см. (19.16)–(19.18)):

$$C = HH^T. \quad (25.23)$$

Тогда априорная ковариация преобразованного вектора  $\tilde{x}$  из (25.20) будет единичной  $n \times n$ -матрицей, а априорно ожидаемое значение  $\tilde{x}$  будет нулевым вектором.

Если для задачи определено значение псевдоранга  $k$ , меньшее  $n$ , и вычисляется решение с минимальной длиной, то разумно такое масштабирование переменных, в результате которого разные компоненты  $\tilde{x}$  будут иметь приблизительно одинаковую неопределенность.

Поскольку числа обусловленности матриц  $A$  и  $\tilde{A}$ , вообще говоря, различны, то можно попытаться выбрать  $H$  так, чтобы уменьшить число обусловленности  $\tilde{A}$ . Если  $A$  не вырождена, то существует матрица  $H$  такая, что  $\text{cond}(AH) = 1$ . Такой матрицей будет, например,  $H = R^{-1}$ , где  $R$  — треугольная матрица, вычисляемая в алгоритме Хаусхолдера. Однако едва ли эта матрица доступна априори. Интересно отметить тем не менее, что если информация, определяющая априорную ковариационную матрицу  $C$  из (25.23), приблизительно соответствует заданным коэффициентам  $[A:b]$ , то матрица  $H$  будет близка к  $R^{-1}$ . В этом случае число  $\text{cond}(AH)$  будет, скорей всего, довольно малым.

Если не имеется оценки для  $C$ , то остается еще возможность взять диагональную матрицу  $H$  как матрицу масштабирования, уравнивающую евклидовы (или любые другие) нормы столбцов  $\tilde{A}$ . Такое уравнивание достигается выбором

$$h_{jj} = \begin{cases} \|a_j\|^{-1}, & a_j \neq 0, \\ 1, & a_j = 0, \end{cases} \quad (25.24)$$

где  $a_j$  —  $j$ -й столбец  $A$ . В [183] доказано, что если  $H$  определить в соответствии с (25.24), где норма взята евклидова, то  $\text{cond}(AH)$  не более чем множителем  $n^{1/2}$  отличается от минимального числа обусловленности, которое можно получить масштабированием столбцов.

Уменьшение числа обусловленности имеет то преимущество, что оценки возмущений типа (9.10) будут давать менее пессимистические результаты. Может оказаться также, что будет установлено значение псевдоранга  $k = n$ , в то время как для исходной задачи было  $k < n$ , что связано с ненужными осложнениями при вычислении решения.

Матрицу преобразования  $H$  можно выбрать так, чтобы  $\tilde{A}$  (или ее подматрица) имела какую-либо особенно удобную форму, например треугольную или диагональную. Так, если вычислено сингулярное разложение  $A$ ,  $A = USV^T$ , то левое умножение  $[A:b]$  на  $U^T$  и правое умножение  $A$  на  $H = V$  преобразуют  $A$  к диагональной матрице  $S$ . Однако эти умножения обычно не выполняют в явном виде; они реализуются параллельно с вычислением сингулярного разложения (см. гл. 18). Систематический анализ сингулярного разложения будет дан в § 6.

Здесь мы обсудим прием, состоящий в замене задачи  $Ax \cong b$  задачей

$$\begin{bmatrix} A \\ F \end{bmatrix} x \cong \begin{bmatrix} b \\ d \end{bmatrix}, \quad (25.25)$$

где  $F$  —  $l \times n$ -матрица,  $d$  —  $l$ -мерный вектор.

Предположим, что для нас предпочтительней решение  $x$ , близкое к известному вектору  $\xi$ . Это предпочтение можно выразить, полагая в (25.25)  $F = I_n$  и  $d = \xi$ . В частности,  $d = \xi = 0$  означает, что предпочтение отдается  $x$  с малой нормой.

Интенсивность предпочтения можно указать, вводя в определение  $F$  и  $d$  масштабирующий множитель (назовем его  $\sigma$ ):

$$F = \sigma^{-1} I_n, \quad d = \sigma^{-1} \xi.$$

Число  $\sigma$  можно рассматривать как оценку величины неопределенности в  $\xi$ . Таким образом, приписывание  $\sigma$  малого значения заставляет решение быть ближе к  $\xi$ .

Развитием этой идеи является использование различных чисел  $\sigma_i$ ,  $i = 1, \dots, n$ , где  $\sigma_i$  — оценка величины неопределенности в  $i$ -й компоненте  $\xi$ . В этом случае полагают  $d_i = \xi_i / \sigma_i$ ,  $i = 1, \dots, n$ , а в качестве  $F$  берут диагональную матрицу с диагональными элементами  $f_{ii} = \sigma_i^{-1}$ .

Наконец, если имеется достаточная априорная статистическая информация относительно ожидаемого значения  $\xi$  решения  $x$ , чтобы можно было построить симметричную положительно определенную  $n \times n$ -матрицу ковариации  $K$  для  $x - \xi$ , то разумно положить

$$F = L^{-1}, \quad (25.26)$$

где  $L$  — нижний треугольный множитель Холецкого для матрицы  $K$  (см. (19.12)–(19.14)), т.е.

$$K = LL^T. \quad (25.27)$$

Кроме того, полагают

$$d = F\xi. \quad (25.28)$$

Отметим, что если система  $Fw = d$  совместна, то случай  $d \neq 0$  простым переносом можно свести к случаю  $d = 0$ . В самом деле, пусть  $w$  — решение системы  $Fw = d$ . Сделаем в (25.25) замену переменных  $x = w + \tilde{x}$ . Она приводит к преобразованной задаче

$$\begin{bmatrix} A \\ F \end{bmatrix} \tilde{x} \cong \begin{bmatrix} b - Aw \\ 0 \end{bmatrix}. \quad (25.29)$$

Если предполагается находить решение с минимальной длиной, то важно произвести эту замену переменных, так как предпочтение малым значениям  $\|\tilde{x}\|$  согласуется с условиями  $F\tilde{x} \cong 0$ .

Рассмотрим теперь вопрос об относительном взвешивании двух систем условий  $Ax \cong b$  и  $Fx \cong d$  в (25.25). С формальной статистической точки зрения можно сказать, что подходящее относительное взвешивание будет

установлено, если представить задачу в виде

$$\begin{bmatrix} G & 0 \\ 0 & F \end{bmatrix} \begin{bmatrix} A \\ I \end{bmatrix} x \cong \begin{bmatrix} G & 0 \\ 0 & F \end{bmatrix} \begin{bmatrix} b \\ \xi \end{bmatrix}, \quad (25.30)$$

где  $(G^T G)^{-1}$  — априорная ковариационная матрица неопределенности в заданном векторе  $b$ , а  $(F^T F)^{-1}$  — априорная ковариационная матрица неопределенности в априорно ожидаемом значении  $\xi$  решения  $x$ . На практике, однако, априорные ковариационные матрицы, особенно  $(F^T F)^{-1}$ , могут быть известны лишь весьма приближенно. Поэтому желательно исследовать влияние изменений в относительных весах на решение и вектор невязки.

С этой целью введем в задачу (25.30) неотрицательный скалярный весовой параметр  $\lambda$  и рассмотрим новую задачу:

$$\begin{bmatrix} \tilde{A} \\ \lambda F \end{bmatrix} x \cong \begin{bmatrix} \tilde{b} \\ \lambda d \end{bmatrix}, \quad (25.31)$$

где

$$\tilde{A} = GA, \quad (25.32)$$

$$\tilde{b} = Gb, \quad (25.33)$$

$$d = F\xi. \quad (25.34)$$

Для читателей, привычных к другим способам мотивировки и формулирования задачи наименьших квадратов с априорными ковариационными матрицами, заметим, что (25.31) равносильно задаче отыскания вектора  $x$ , минимизирующего квадратичную функцию

$$\|\tilde{A}x - \tilde{b}\|^2 + \lambda^2 \|Fx - d\|^2,$$

или, что одно и то же,

$$(Ax - b)^T (G^T G) (Ax - b) + \lambda^2 (x - \xi)^T (F^T F) (x - \xi).$$

Идея использовать относительный весовой параметр  $\lambda$  в этом контексте была высказана в статье [117]. Усовершенствования и приложения этой техники указаны в [121, 122, 130, 96–99]. После статьи Марквардта [121] и программы, представленной им в SHARE (организация по обмену вычислительной информацией), применение рассматриваемого подхода к линейризации нелинейных задач наименьших квадратов часто называют *методом Марквардта*. Для исследования задачи (25.31) в зависимости от  $\lambda$  иногда используют еще термин *гребневая регрессия*.

Чтобы проанализировать зависимость решения и невязки в задаче (25.31) от параметра  $\lambda$ , произведем прежде всего замену переменных

$$x = \xi + F^{-1}y, \quad (25.35)$$

что приводит к преобразованной задаче:

$$\begin{bmatrix} \hat{A} \\ \lambda I \end{bmatrix} y \cong \begin{bmatrix} \hat{b} \\ 0 \end{bmatrix}, \quad (25.36)$$

где  $\hat{A} = \tilde{A}F^{-1}$  и  $\hat{b} = \tilde{b} - \tilde{A}\xi$ .

масштабирование посредством  $\Gamma^{-1}$ , использованные в (25.35), обсуждались в § 3. Их цель — перейти к новой переменной  $y$ , масштабированной лучше, чем  $x$ , что позволяет получить более осмысленную оценку вектора решения.

Запишем сингулярное разложение матрицы  $\hat{A}$ :

$$\hat{A}_{m \times n} = U_{m \times m} \begin{bmatrix} S_{n \times n} \\ 0_{(m-n) \times n} \end{bmatrix} V_{n \times n}^T. \quad (25.37)$$

Напомним, что  $S = \text{Diag} \{s_1, \dots, s_n\}$ . Если  $\text{rank } A = k < n$ , то  $s_i = 0$  для  $i > k$ . Выполним ортогональную замену переменных

$$y = Vp \quad (25.38)$$

и умножим (25.36) слева на ортогональную матрицу

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix}.$$

В результате получим новую задачу наименьших квадратов:

$$\begin{bmatrix} S_{n \times n} \\ 0_{(m-n) \times n} \\ \lambda I_n \end{bmatrix} p \cong \begin{bmatrix} g \\ 0 \end{bmatrix} \quad (25.39)$$

где

$$g = U^T \hat{b}. \quad (25.40)$$

Наконец, при  $\lambda > 0$  из задачи (25.39) можно исключить подматрицу  $\lambda I_n$  посредством левого умножения на соответствующие вращения Гивенса. Так, чтобы исключить  $i$ -й диагональный элемент  $\lambda I_n$ , умножим (25.39) на матрицу, которая отличается от единичной матрицы порядка  $m+n$  только следующими четырьмя позициями:

|              |                                              |                                             |               |
|--------------|----------------------------------------------|---------------------------------------------|---------------|
|              | Столбец $i$                                  |                                             | Столбец $m+i$ |
|              | $\vdots$                                     |                                             | $\vdots$      |
| Строка $i$   | $\frac{s_i}{(s_i^2 + \lambda^2)^{1/2}}$      | $\frac{\lambda}{(s_i^2 + \lambda^2)^{1/2}}$ | $\vdots$      |
|              | $\vdots$                                     | $\vdots$                                    | $\vdots$      |
| Строка $m+i$ | $-\frac{\lambda}{(s_i^2 + \lambda^2)^{1/2}}$ | $\frac{s_i}{(s_i^2 + \lambda^2)^{1/2}}$     | $\vdots$      |
|              | $\vdots$                                     | $\vdots$                                    | $\vdots$      |

После того как это исключение проделано для  $i = 1, \dots, n$ , будет получена эквивалентная задача:

$$\begin{bmatrix} S_{n \times n}^{(\lambda)} \\ 0_{(m-n) \times n} \\ 0_{n \times n} \end{bmatrix} p \cong \begin{bmatrix} g^{(\lambda)} \\ h^{(\lambda)} \end{bmatrix} \quad (25.41)$$

здесь

$$g_i^{(\lambda)} = \begin{cases} \frac{g_i s_i}{(s_i^2 + \lambda^2)^{1/2}}, & i = 1, \dots, n \\ g_i, & i = n+1, \dots, m, \end{cases} \quad (25.42)$$

$$h_i^{(\lambda)} = -\frac{g_i \lambda}{(s_i^2 + \lambda^2)^{1/2}}, \quad i = 1, \dots, n, \quad (25.43)$$

$$S^{(\lambda)} = \text{diag} \{s_1^{(\lambda)}, \dots, s_n^{(\lambda)}\},$$

где

$$s_i^{(\lambda)} = (s_i^2 + \lambda^2)^{1/2}, \quad i = 1, \dots, n. \quad (25.44)$$

При  $\lambda = 0$  в задаче (25.39) матрица диагональна ранга  $k$ , а решение  $p^{(0)}$  имеет компоненты

$$p_i^{(0)} = \begin{cases} \frac{g_i}{s_i}, & i = 1, \dots, k, \\ 0, & i = k+1, \dots, n. \end{cases} \quad (25.45)$$

При  $\lambda > 0$  для компонент решения получаем из (25.41)–(25.44) выражения

$$p_i^{(\lambda)} = \begin{cases} \frac{g_i^{(\lambda)}}{s_i^{(\lambda)}} = \frac{g_i s_i}{s_i^2 + \lambda^2} = p_i^{(0)} \frac{s_i^2}{s_i^2 + \lambda^2}, & i = 1, \dots, k, \\ 0, & i = k+1, \dots, n. \end{cases} \quad (25.46)$$

Кроме того,

$$\|p^{(\lambda)}\|^2 = \sum_{i=1}^k \left[ \frac{g_i s_i}{s_i^2 + \lambda^2} \right]^2 = \sum_{i=1}^k [p_i^{(0)}]^2 \left[ \frac{s_i^2}{s_i^2 + \lambda^2} \right]^2. \quad (25.47)$$

Отметим, что формулы (25.46) и (25.47) остаются справедливыми и при  $\lambda = 0$ .

Вектор  $p^{(\lambda)}$  — решение задачи (25.39) — можно преобразовать, пользуясь линейными соотношениями (25.38) и (25.35), в векторы  $y^{(\lambda)}$  и  $x^{(\lambda)}$ , которые решают соответственно задачи (25.36) и (25.31).

Нас интересует главным образом задача  $\tilde{A}x \cong \tilde{b}$ . Задача (25.31) была введена лишь как техническое средство генерирования решений для конкретного семейства родственных задач. Поэтому выведем еще формулы для величины  $\omega_\lambda = \|\tilde{b} - \tilde{A}x^{(\lambda)}\|$ . Для  $\lambda > 0$  и  $p^{(\lambda)}$ , определенного в (25.46), имеем

$$\begin{aligned} \omega_\lambda^2 &= \|\tilde{b} - \tilde{A}x^{(\lambda)}\|^2 = \|\hat{b} - \hat{A}y^{(\lambda)}\|^2 = \left\| g - \begin{bmatrix} S \\ 0 \end{bmatrix} p^{(\lambda)} \right\|^2 = \\ &= \sum_{i=1}^k [g_i - s_i p_i^{(\lambda)}]^2 + \sum_{i=k+1}^m s_i^2 = \sum_{i=1}^k g_i^2 \left[ \frac{\lambda^2}{s_i^2 + \lambda^2} \right]^2 + \sum_{i=k+1}^m s_i^2. \end{aligned} \quad (25.48)$$

Заметим, что при увеличении  $\lambda$  происходит уменьшение  $\|p^{(\lambda)}\|$  и увеличение  $\omega_\lambda$ . Поэтому хозяин задачи имеет возможность выбрать  $\lambda$ , обеспечивающее приемлемый компромисс между величиной решения  $p^{(\lambda)}$  и значением нормы невязки  $\omega_\lambda$ .

Получаемое в описанном методе множество решений задачи (25.31) имеет важное свойство оптимальности, выражаемое следующей теоремой.

**Теорема 25.49** [121, 130]. Пусть фиксировано неотрицательное значение  $\bar{\lambda}$  параметра  $\lambda$ , и пусть  $\bar{y}$  — соответствующее решение задачи (25.36). Положим  $\bar{\omega} = \|\hat{b} - \hat{A}\bar{y}\|$ . Тогда  $\bar{\omega}$  будет минимальным значением  $\|\hat{b} - \hat{A}y\|$  на множестве векторов  $y$ , для которых  $\|y\| \leq \|\bar{y}\|$ .

**Доказательство.** Предположим противное. Тогда найдется вектор  $\tilde{y}$ , для которого  $\|\tilde{y}\| \leq \|\bar{y}\|$  и  $\|\hat{b} - \hat{A}\tilde{y}\| < \|\hat{b} - \hat{A}\bar{y}\|$ . Следовательно,  $\|\hat{b} - \hat{A}\tilde{y}\|^2 + \bar{\lambda}^2 \|\tilde{y}\|^2 < \|\hat{b} - \hat{A}\bar{y}\|^2 + \bar{\lambda}^2 \|\bar{y}\|^2$ , а это противоречит предположению, что  $\bar{y}$  есть решение задачи (25.36) и потому минимизирует  $\|\hat{b} - \hat{A}y\|^2 + \bar{\lambda}^2 \|y\|^2$ .

При исследовании конкретной задачи наименьших квадратов очень полезным может быть простое табличное или графическое изображение величин  $\|p^{(\lambda)}\|$  и  $\omega_\lambda$  из формул (25.47), (25.48). Дополнительную информацию может дать табличное или графическое изображение отдельных компонент решения (при этом нужно опираться на формулы (25.46), (25.38) и (25.35)) как функций от  $\lambda$ . Пример анализа этого типа дан в гл. 26.

Значение  $\lambda$ , обеспечивающее предписанную норму решения или норму невязки, можно найти, решая уравнения (25.47) или (25.48). Если для этой цели используется метод Ньютона, то могут пригодиться следующие выражения:

$$t = \lambda^2, \quad \varphi_i(t) = \frac{g_i s_i}{s_i^2 + t},$$

$$\|p^{(\lambda)}\|^2 = \sum_{i=1}^k \varphi_i^2(t), \quad \frac{d(\|p^{(\lambda)}\|^2)}{dt} = -2 \sum_{i=1}^k \frac{\varphi_i^2(t)}{s_i^2 + t};$$

$$u = \lambda^{-2}, \quad \psi_i(u) = \frac{g_i}{u s_i^2 + 1},$$

$$\omega_\lambda^2 = \sum_{i=1}^k \psi_i^2(u) + \sum_{i=k+1}^m g_i^2, \quad \frac{d(\omega_\lambda^2)}{du} = -2 \sum_{i=1}^k \frac{\psi_i^2(u) s_i^2}{u s_i^2 + 1}.$$

Если задача (25.36) решается непосредственно \*) с помощью какого-либо алгоритма, определяющего норму  $\rho_\lambda$  вектора невязки, то следует обратить внимание на соотношение

$$\rho_\lambda^2 = \omega_\lambda^2 + \lambda^2 \|y^{(\lambda)}\|^2. \quad (25.50)$$

Если нужно найти величину  $\omega_\lambda$ , то можно вычислить  $\|y^{(\lambda)}\|^2$ , а затем разрешить (25.50) относительно  $\omega_\lambda$ .

\*) То есть без преобразования к (25.39), а затем к (25.41). (Примеч. пер.)

## § 5. Удаление переменных

Удаление переменной из задачи эквивалентно фиксированию значения этой переменной в нуле. Если удаляется одна переменная, скажем  $x_n$ , то задача

$$Ax \cong b \quad (25.51)$$

преобразуется в

$$\tilde{A}\tilde{x} \cong b, \quad (25.52)$$

где  $\tilde{A}$  —  $m \times (n-1)$ -матрица, образованная первыми  $n-1$  столбцами  $A$ , а  $\tilde{x}$  —  $n-1$ -мерный вектор.

Предположим, что  $m \geq n$ . Согласно теореме 5.12, сингулярные числа  $\tilde{A}$  разделяют сингулярные числа  $A$ , откуда следует, что  $\text{cond}(\tilde{A}) \leq \text{cond}(A)$ . Повторное применение этого процесса показывает, что удаление любого собственного подмножества переменных приводит к матрице с числом обусловленности не большим, чем у исходной матрицы.

Ясно, что минимальная невязка в задаче (25.52) не меньше, чем минимальная невязка задачи (25.51).

Таким образом, удаление переменных, как и некоторые из ранее обсуждавшихся приемов, является средством снижения (или по крайней мере неувеличения) числа обусловленности матрицы ценой увеличения (или по крайней мере неуменьшения) нормы вектора невязки.

Мы сочли полезным отметить эти свойства, с тем чтобы сопоставить удаление переменных или введение дополнительных переменных с другими методами стабилизации. Тем не менее не они (свойства) обычно являются мотивировкой для процедур этого типа. Чаше к удалению или введению переменных прибегают для того, чтобы определить наименьшее число параметров, которые можно использовать в решении, сохраняя приемлемо малую норму невязки.

В некоторых случаях имеется естественное упорядочение переменных; примером могут служить коэффициенты многочлена от одного неизвестного. В таких случаях очевидным образом получается последовательность решений: сначала берется только первая переменная, затем первая и вторая и т.д. Для задачи сглаживания посредством многочленов или отрезков ряда Фурье разработаны весьма специализированные алгоритмы (см., например, [55]).

Если нет естественного упорядочения переменных, то иногда рассматривают следующую задачу о выборе подмножества. Для каждого значения  $k = 1, \dots, n$  нужно найти подмножество  $J_k$  из  $k$  индексов такое, что норма  $\rho_k$  невязки, полученной при решении задачи только относительно  $k$  переменных  $x_i, i \in J_k$ , не превосходит значения, получаемого при выборе любого другого набора из  $k$  переменных. Обычно в задачу о выборе вводят еще допуск  $\tau$  на линейную независимость и сравнивают лишь такие множества из  $k$  индексов, что ассоциированные с ними матрицы удовлетворяют некоторому тесту на обусловленность, связанному с  $\tau$  (например, проверяют величины главных элементов).

Очевидно, что последовательность  $\{\rho_k\}$  не возрастает с увеличением  $k$ . Поэтому можно установить границу  $\bar{\rho}$  с таким расчетом, что процесс за-



канчивается, когда достигнуто значение  $k$ , для которого  $\rho_k < \bar{\rho}$ . Другая возможность — остановить вычисления, когда значение  $\rho_k - \rho_{k+1}$  становится меньше назначенной границы, которая может зависеть от  $k$ . К такому тесту на окончание приводят статистические соображения, связанные с  $F$ -распределением (см., например, [144]).

Задача о выборе в том виде, как она сформулирована, требует изнурительной работы по перебору всех  $\binom{n}{k}$  комбинаций из  $k$  переменных. При больших значениях  $\binom{n}{k}$  такой подход становится чересчур дорогостоящим.

Были разработаны методы, использующие для определения оптимальных подмножеств  $J_k$  частичное упорядочение подмножеств (по отношению включения); при этом не требуется явной проверки каждого подмножества. Изложение этих идей можно найти в работах [83, 113] и в упомянутых там работах. Предложение основать выбор подмножества на гребневой регрессии высказано в [99].

Альтернативным образом действий является компромисс, при котором рассматривают следующую, более ограниченную задачу. Вначале решают при  $k = 1$  прежнюю задачу. Пусть  $\tilde{J}_1$  — ее решение. После того как определено множество  $\tilde{J}_k$ , сравниваются в поиске предпочтительного множества из  $k + 1$  переменных лишь множества вида  $\tilde{J}_k \cup \{j\}$ , где  $j \notin \tilde{J}_k$ ; выбранное множество обозначается через  $\tilde{J}_{k+1}$ . Пусть последовательности найденных этим путем множеств  $\tilde{J}_k, k = 1, \dots, n$ , соответствует последовательность норм невязок  $\tilde{\rho}_k$ . Заметим, что  $\tilde{J}_k = J_k$  и  $\tilde{\rho}_k = \rho_k$  при  $k = 1$  и  $k = n$ , но для  $1 < k < n$  множество  $\tilde{J}_k$ , вообще говоря, отличается от  $J_k$ , а  $\tilde{\rho}_k > \rho_k$ .

Алгоритм этого типа называют пошаговой регрессией (см. [151, с. 191–203]). Вычисления можно организовать так, что выбор новой переменной на шаге  $k$  будет лишь немногим сложнее, чем алгоритм выбора главного элемента, обычно используемый при решении системы линейных уравнений. Эту идею можно усовершенствовать, рассматривая на каждом шаге возможность удаления переменных, чей вклад в уменьшение невязки стал незначительным в результате введения новых переменных [151].

Основное математическое затруднение в пошаговой регрессии состоит в том, что величина вклада, вносимого отдельной переменной в уменьшение нормы невязки, в общем случае зависит от того, какие переменные включаются вместе с ней в решение. Эту трудность можно обойти, выполняя линейный переход к новым переменным, чьи индивидуальные воздействия на вектор невязки взаимно независимы. В статистической терминологии новые переменные не коррелированы. На алгебраическом языке этот переход равносильен замене матрицы  $A$  новой матрицей  $\tilde{A} = AC$ , у которой столбцы ортогональны. Имеется множество различных матриц  $C$ , реализующих такое преобразование, и, вообще говоря, различные матрицы порождают различные наборы некоррелированных переменных. Одной из трансформирующих матриц, имеющей вдобавок другие полезные свойства, является матрица  $V$  из сингулярного разложения  $A = USV^T$  матрицы  $A$ .

Если, в частности, требуется, чтобы трансформирующая матрица  $\tilde{C}$  была ортогональна, то, чтобы  $AC$  имела ортогональные столбцы,  $C$  обязательно должна совпадать с матрицей  $V$  некоторого сингулярного разложения  $A$ .

Добавим, что числа  $t_i$  — квадратные корни из диагональных элементов диагональной матрицы  $[(AC)^T(AC)]^{-1}$  — имеют следующую статистическую интерпретацию: с точностью до общего множителя они совпадают со среднеквадратичными отклонениями новых переменных. При выборе  $C = V$  эти числа  $t_i$  становятся обратными к сингулярным числам матрицы  $A$ . Этот выбор минимизирует наименьшее из среднеквадратичных отклонений, а также каждое из отклонений в предположении, что предыдущие переменные уже выбраны.

Обсуждение сингулярного разложения матрицы  $A$  будет продолжено в следующем параграфе.

## § 6. Сингулярный анализ

Пусть для матрицы  $A$  найдено сингулярное разложение (см. гл. 4, 18)

$$A = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^T. \quad (25.53)$$

$$\text{Тогда можно вычислить вектор} \quad g = U^T b \quad (25.54)$$

$$\text{и рассмотреть задачу наименьших квадратов} \quad \begin{bmatrix} S \\ 0 \end{bmatrix} p \cong g, \quad (25.55)$$

$$\text{где } p \text{ связано с } x \text{ ортогональным линейным преобразованием} \quad x = Vp. \quad (25.56)$$

Задача (25.55) эквивалентна задаче  $Ax \cong b$  в том смысле, как эта эквивалентность была определена в гл. 2 для произвольных ортогональных преобразований задачи наименьших квадратов.

Поскольку  $S$  — диагональная матрица ( $S = \text{Diag}\{s_1, \dots, s_n\}$ ), то влияние каждой компоненты  $p$  на норму невязки видно непосредственно. Вводя в решение компоненту  $p_i$  со значением

$$\hat{p}_i = g_i/s_i, \quad (25.57)$$

мы снижаем квадрат нормы невязки на величину  $g_i^2$ .

Предположим, что сингулярные числа упорядочены так, что  $s_k \geq s_{k+1}$ ,  $k = 1, \dots, n-1$ . Тогда естественно рассмотреть "пробные" решения задачи (25.55) следующего вида:

$$p^{(k)} = \begin{bmatrix} \hat{p}_1 \\ \vdots \\ \hat{p}_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad k = 0, 1, \dots, n. \quad (25.58)$$



числа  $p_j$  определяются формулами (25.57). Пробный вектор  $p^{(k)}$  является нормальным псевдорешением задачи (25.55), если сингулярные числа  $s_j, j > k$ , считаются нулевыми.

Из пробных векторов  $p^{(k)}$  получаем пробные решения  $x^{(k)}$  задачи  $Ax \cong b$ :

$$x^{(k)} = Vp^{(k)} = \sum_{j=1}^k \hat{p}_j v^{(j)}, \quad k = 0, \dots, n. \quad (25.59)$$

Здесь  $v^{(j)}$  —  $j$ -й столбец  $V$ . Заметим, что

$$\|x^{(k)}\|^2 = \|p^{(k)}\|^2 = \sum_{j=1}^k \hat{p}_j^2 = \sum_{j=1}^k \left( \frac{g_j}{s_j} \right)^2, \quad (25.60)$$

и, следовательно,  $\|x^{(k)}\|$  — неубывающая функция от  $k$ . Квадрат нормы невязки, отвечающей вектору  $x^{(k)}$ , равен

$$\rho_k^2 = \|b - Ax^{(k)}\|^2 = \sum_{j=k+1}^m g_j^2. \quad (25.61)$$

Исследование столбцов матрицы  $V$ , ассоциированных с малыми сингулярными числами, — очень эффективное средство выявления почти линейно зависимых наборов столбцов  $A$  (см. (12.23), (12.24)).

Наша практика показывает, что вычисление и вывод на печать матрицы  $V$  и величин  $s_k, s_k^{-1}, \hat{p}_k, x^{(k)}, \|x^{(k)}\|, g_k, g_k^2, \rho_k^2$  и  $\rho_k$  для всех значений  $k$  (от 0 или 1 до  $n$ ) чрезвычайно полезны при анализе трудных практических задач метода наименьших квадратов. Для некоторых статистических интерпретаций (см. (12.2)) интерес представляют также величины

$$\sigma_k = \left[ \frac{\rho_k^2}{m - k} \right]^{1/2}, \quad k = 0, \dots, n. \quad (25.62)$$

Предположим, что матрица  $A$  плохо обусловлена. Тогда некоторые из сингулярных чисел с большими номерами будут существенно меньше предшествующих. В этом случае некоторые значения  $\hat{p}_j$  с большими номерами могут быть слишком велики. В типичной ситуации пытаются найти такой индекс  $k$ , чтобы все коэффициенты  $p_j, j \leq k$ , были достаточно малы, все сингулярные числа  $s_j, j \leq k$ , достаточно велики, а норма невязки  $\rho_k$  достаточно мала. Если такой индекс  $k$  существует, то в качестве приемлемого решения можно взять пробный вектор  $x^{(k)}$ .

Эта техника успешно применялась во многих приложениях (см., например, [87], где рассматривается приложение к численному решению фредгольмовых интегральных уравнений первого рода).

Как только вычислены сингулярные числа и вектор  $g$ , уже просто вычислить для ряда значений  $\lambda$  числа  $\|p^{(\lambda)}\|$  и  $\omega_\lambda$ , нужные в методе стабилизации Левенберга—Марквардта (см. (25.47) и (25.48)). Эти величины представляют интерес из-за свойства оптимальности, выражаемого теоремой 25.49.

В следующей главе рассматривается пример, в котором все упомянутые здесь величины вычисляются и интерпретируются для конкретного набора входных данных.

## ПРИМЕРЫ НЕКОТОРЫХ МЕТОДОВ АНАЛИЗА ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ

Рассмотрим задачу наименьших квадратов  $Ax \cong b$ , матрица коэффициентов которой приведена в табл. 26.1. Мы предположим, что имеется неопределенность порядка  $0,5 \times 10^{-8}$  в элементах  $A$  и порядка  $0,5 \times 10^{-4}$  в элементах  $b$ .

Рассмотрим вначале сингулярный анализ этой задачи, как он описан в § 6 гл. 25. Результаты, полученные при выполнении программы сингулярного анализа на машине UNIVAC 1108, воспроизведены на рис. 26.1.

В общем случае программа производит замену переменных  $x = Du$  и решает задачу  $(AD)u \cong b$ . В данном примере мы положили  $D = I$ , поэтому символ  $u$  на рис. 26.1 нужно отождествить с  $x$ .

Напомним, что вектор  $g$  (столбец выдачи, стоящий под шапкой "G") вычисляется по формуле  $g = U^T b$ , где  $U$  — ортогональная матрица. Так как четвертая и пятая компоненты  $g$  меньше предполагаемой неопределенности в  $b$ , то мы хотели бы рассматривать эти две компоненты как нулевые. Это свидетельствует в пользу того, чтобы считать третий пробный вектор  $x^{(3)}$  (столбец выдачи под шапкой "SOLN 3") наиболее удовлетворительным решением.

Другой метод состоит в сравнении чисел  $\sigma_k$  (см. (25.62)), стоящих в столбце под шапкой "N.S.R.C.S.S." (что является сокращением от Normalized Square Root of Cumulative Sum of Squares, т.е. нормированный квадратный корень из накопленной суммы квадратов), с предполагаемой неопределенностью  $0,5 \times 10^{-4}$  в  $b$ . Замечаем, что  $\sigma_2 (= 1,1107 \times 10^{-2})$  значительно больше этой предполагаемой неопределенности, в то время как  $\sigma_3 (= 4,0548 \times 10^{-5})$  чуть меньше ее. Это может служить основанием для того, чтобы выбрать  $x^{(3)}$  в качестве предпочтительного пробного решения.

Если предполагаемая неопределенность в  $b$  интерпретируется статистически как среднеквадратичное отклонение погрешностей в  $b$ , то произведение этой величины ( $0,5 \times 10^{-4}$ ) на обратное к сингулярному числу (эти обратные числа находятся в столбце с шапкой "1/синг.число") дает среднеквадратичное отклонение соответствующей компоненты вектора  $p$  (столбец с шапкой "P"). Как видим, первые три компоненты  $p$  превосходят по абсолютной величине соответствующие среднеквадратичные отклонения, в то время как последние две компоненты меньше своих отклонений. Это также может быть причиной, чтобы предпочесть пробное решение  $x^{(3)}$ .

Есть и еще один метод выбора  $x^{(k)}$ . Пробные решения  $x^{(1)}$  и  $x^{(2)}$  будут, по всей вероятности, отклонены из-за того, что для них нормы невязок слишком велики (см. столбец, озаглавленный "RNORM"). Вектор  $x^{(5)}$  тоже, по-видимому, будет отвергнут, поскольку  $\|x^{(5)}\|$  слишком велика (см. столбец, озаглавленный "YNORM"). Выбор между  $x^{(3)}$  и  $x^{(4)}$  не так очевиден. Все же, вероятно,  $x^{(3)}$  следует предпочесть вектору  $x^{(4)}$ .

Таблица 26.1

Матрица исходных данных [A : b]

|             |             |             |             |             |         |
|-------------|-------------|-------------|-------------|-------------|---------|
| -0,13405547 | -0,20162827 | -0,16930778 | -0,18971990 | -0,17387234 | -0,4361 |
| -0,10379475 | -0,15766336 | -0,13346256 | -0,14948550 | -0,13597690 | -0,3437 |
| -0,08779597 | -0,12883867 | -0,10683007 | -0,12011796 | -0,10932972 | -0,2657 |
| 0,02058554  | 0,00335331  | -0,01641270 | 0,00078606  | 0,00271659  | -0,0392 |
| -0,03248093 | -0,01876799 | 0,00410639  | -0,01405894 | -0,01384391 | 0,0193  |
| 0,05967662  | 0,06667714  | 0,04352153  | 0,05740438  | 0,05024962  | 0,0747  |
| 0,06712457  | 0,07352437  | 0,04489770  | 0,06471862  | 0,05876455  | 0,0935  |
| 0,08687186  | 0,09368296  | 0,05672327  | 0,08141043  | 0,07302320  | 0,1079  |
| 0,02149662  | 0,06222662  | 0,07213486  | 0,06200069  | 0,05570931  | 0,1930  |
| 0,06687407  | 0,10344506  | 0,09153849  | 0,09508223  | 0,08393667  | 0,2058  |
| 0,15879069  | 0,18088339  | 0,11540692  | 0,16160727  | 0,14796479  | 0,2606  |
| 0,17642887  | 0,20261830  | 0,13057860  | 0,18385729  | 0,17005549  | 0,3142  |
| 0,11414080  | 0,17259611  | 0,14816471  | 0,16007466  | 0,14374096  | 0,3529  |
| 0,07846038  | 0,14669563  | 0,14365800  | 0,14003842  | 0,12571177  | 0,3615  |
| 0,10803175  | 0,16994623  | 0,14971519  | 0,15885312  | 0,14301547  | 0,3647  |

СИНГУЛЯРНЫЙ АНАЛИЗ ЗАДАЧИ НАИМЕНЬШИХ КВАДРАТОВ  $AX = B$ , МАСШТАБИРОВАННОЙ КАК  $(AD)Y = B$ 

$$M = 15, N = 5, MDATA = 15$$

Вариант масштабирования № 1. D — единичная матрица.

V — матрица из сингулярного разложения матрицы AD (элементы V умножены на  $10^4$ ).

| Индекс | Синг. число  | P            | 1/синг. число | G            | G <sup>2</sup> | C.S.S.      | N.S.R.C.S.S. |
|--------|--------------|--------------|---------------|--------------|----------------|-------------|--------------|
| 0      |              |              |               |              |                |             |              |
| 1      | 1,0000       | 9,9981 - 01  | 1,0000 + 00   | 9,9981 - 01  | 9,9963 - 01    | 1,0412 + 00 | 2,6347 - 01  |
| 2      | 0,1000       | 2,0003 + 00  | 1,0000 + 01   | 2,0003 - 01  | 4,0013 - 02    | 4,1617 - 02 | 5,4522 - 02  |
| 3      | 0,0100       | -4,0047 + 00 | 1,0000 + 02   | -4,0047 - 02 | 1,6038 - 03    | 1,6038 - 03 | 1,1107 - 02  |
| 4      | 0,9997 - 05  | 1,7755 + 00  | 1,0003 + 05   | 1,7750 - 05  | 3,1506 - 10    | 1,9730 - 08 | 4,0548 - 05  |
| 5      | 0,9904 - 07  | 1,6761 + 02  | 1,0097 + 07   | 1,6599 - 05  | 2,7554 - 10    | 1,9415 - 08 | 4,2012 - 05  |
|        |              |              |               |              |                | 1,9139 - 08 | 4,3748 - 05  |
| Индекс | YNORM        | RNORM        | lg (YNORM)    | lg (RNORM)   |                |             |              |
| 0      | 0,00000      | 0,10204 + 01 | -1,0000,00000 | 0,00878      |                |             |              |
| 1      | 0,99981 + 00 | 0,20400 + 00 | -0,00008      | -0,69036     |                |             |              |
| 2      | 0,22363 + 01 | 0,40047 - 01 | 0,34953       | -1,39743     |                |             |              |
| 3      | 0,45868 + 01 | 0,14046 - 03 | 0,66151       | -3,85244     |                |             |              |
| 4      | 0,49184 + 01 | 0,13934 - 03 | 0,69183       | -3,85593     |                |             |              |
| 5      | 0,16768 + 03 | 0,13834 - 03 | 2,22449       | -3,85904     |                |             |              |

Рис. 26.1. Выводы подпрограммы SVA для иллюстративной задачи гл. 26

НОРМЫ РЕШЕНИЯ И НЕВЯЗКИ ДЛЯ РЯДА ЗНАЧЕНИЙ ПАРАМЕТРА  $\lambda$  В МЕТОДЕ ЛЕВЕНБЕРГА-МАРКВАРДТА

| LAMBDA       | YNORM        | RNORM        | lg (LAMBDA) | lg (YNORM) | lg (RNORM) |
|--------------|--------------|--------------|-------------|------------|------------|
| 0,10000 + 02 | 0,99012 - 02 | 0,10107 + 01 | 1,00000     | -2,00431   | 0,00463    |
| 0,35464 + 01 | 0,73657 - 01 | 0,94834 + 00 | 0,54979     | -1,13279   | -0,02303   |
| 0,12577 + 01 | 0,38746 + 00 | 0,64525 + 00 | 0,09958     | -0,41178   | -0,19027   |
| 0,44603 + 00 | 0,83939 + 00 | 0,25574 + 00 | -0,35063    | -0,07604   | -0,59221   |
| 0,15818 + 00 | 0,11304 + 01 | 0,15037 + 00 | -0,80084    | 0,05325    | -0,82283   |
| 0,56098 - 01 | 0,18231 + 01 | 0,61718 - 01 | -1,25105    | 0,26080    | -1,20959   |
| 0,19895 - 01 | 0,23138 + 01 | 0,32867 - 01 | -1,70126    | 0,36433    | -1,48324   |
| 0,70555 - 02 | 0,34800 + 01 | 0,13348 - 01 | -2,15147    | 0,54158    | -1,87460   |
| 0,25022 - 02 | 0,43817 + 01 | 0,23671 - 02 | -2,60168    | 0,64165    | -2,62579   |
| 0,88737 - 03 | 0,45594 + 01 | 0,34333 - 03 | -3,05189    | 0,65891    | -3,46429   |
| 0,31470 - 03 | 0,45833 + 01 | 0,14596 - 03 | -3,50210    | 0,66118    | -3,83578   |
| 0,11161 - 03 | 0,45864 + 01 | 0,14053 - 03 | -3,95231    | 0,66147    | -3,85222   |
| 0,39580 - 04 | 0,45880 + 01 | 0,14033 - 03 | -4,40252    | 0,66162    | -3,85284   |
| 0,14037 - 04 | 0,46256 + 01 | 0,13983 - 03 | -4,85273    | 0,66516    | -3,85439   |
| 0,49780 - 05 | 0,48028 + 01 | 0,13938 - 03 | -5,30295    | 0,68150    | -3,85580   |
| 0,17654 - 05 | 0,49274 + 01 | 0,13933 - 03 | -5,75316    | 0,69262    | -3,85595   |
| 0,62609 - 06 | 0,63958 + 01 | 0,13929 - 03 | -6,20337    | 0,80590    | -3,85608   |
| 0,22204 - 06 | 0,28244 + 02 | 0,13904 - 03 | -6,65358    | 1,45092    | -3,85687   |
| 0,78743 - 07 | 0,10281 + 03 | 0,13849 - 03 | -7,10379    | 2,01203    | -3,85857   |
| 0,27926 - 07 | 0,15534 + 03 | 0,13835 - 03 | -7,55400    | 2,19129    | -3,85902   |
| 0,99036 - 08 | 0,16602 + 03 | 0,13834 - 03 | -8,00421    | 2,22017    | -3,85904   |

ПОСЛЕДОВАТЕЛЬНОСТЬ ПРОБНЫХ РЕШЕНИЙ  $x$ 

|   | SOLN 1          | SOLN 2           | SOLN 3           | SOLN 4           | SOLN 5           |
|---|-----------------|------------------|------------------|------------------|------------------|
| 1 | 0,37409643 + 00 | -0,11313896 + 01 | -0,24857328 + 01 | -0,28373911 + 01 | -0,65543824 + 02 |
| 2 | 0,51951898 + 00 | 0,39226783 + 00  | -0,52913252 + 00 | 0,59819891 + 00  | 0,87667385 + 02  |
| 3 | 0,41223419 + 00 | 0,17145393 + 01  | -0,18414114 + 00 | -0,37354642 + 00 | -0,69476752 + 02 |
| 4 | 0,47949432 + 00 | 0,61736754 + 00  | 0,16156794 + 01  | 0,39464945 + 00  | 0,80802341 + 02  |
| 5 | 0,43580915 + 00 | 0,49625358 + 00  | 0,34547871 + 01  | 0,39353651 + 01  | -0,69133711 + 02 |

Рис. 26.1 (окончание)

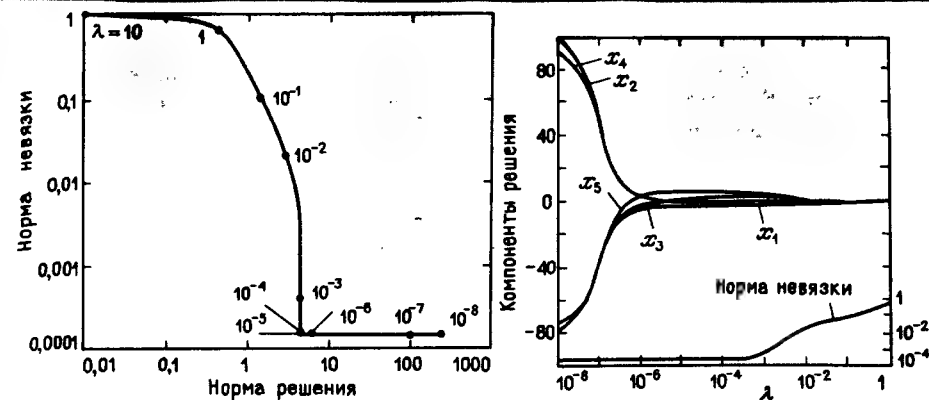
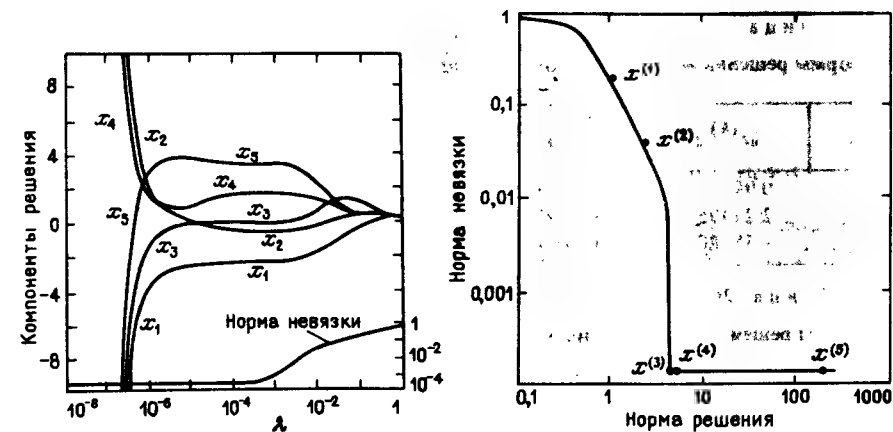
Рис. 26.2. Нормы невязок и нормы решений для ряда значений параметра стабилизации  $\lambda$  в методе Левенберга-МарквардтаРис. 26.3. Компоненты решения и норма невязки как функции от  $\lambda$ .

Рис. 26.4. Те же данные, что и на рис. 26.3, при уменьшенном масштабе по вертикальной оси

Рис. 26.5. Пробные решения  $x^{(i)}$ , полученные в сингулярном анализе, в сопоставлении с континуумом решений Левенберга-Марквардта

так как  $\|x^{(3)}\| < \|x^{(4)}\|$ , а норма невязки, соответствующей  $x^{(4)}$ , лишь незначительно меньше нормы невязки для  $x^{(3)}$ .

Хотя все четыре метода выбора предпочтительного решения  $x^{(k)}$  привели нас к одному и тому же вектору  $x^{(3)}$ , так будет не для любых исходных данных. Пользователь должен решить, какой из этих критериев (а возможно, и какой-то иной) наиболее подходит для его задачи.

Интересно сравнить эти результаты с результатами, которые дают для той же задачи другие методы стабилизации решений. Анализ Левенберга-Марквардта (гл. 25, § 4) приводит к континууму пробных решений. На рис. 26.1 показана информация, нужная для применения метода

Левенберга—Марквардта. С ее помощью можно вычертить график (рис. 26.2), демонстрирующий зависимость между  $RNORM$  и  $YNORM$ . Согласно теореме 25.49, эта кривая будет на плоскости переменных  $(YNORM - RNORM)$  граничной линией для области, соответствующей нашей задаче: для любого вектора  $y$  точка с координатами  $(\|y\|, \|b - Ay\|)$  лежит на кривой или выше ее.

Более детальную информацию дает вычисление в соответствии с (25.46), (25.38) и (25.35) и составление графиков отдельных компонент решения как функций от  $\lambda$ . Рис. 26.3, 26.4 показывают такие графики для данного примера. Этот тип графиков подробно обсуждается в [99].

На рис. 26.5 дано сопоставление норм решений и невязок для пяти пробных решений, полученных методом сингулярного анализа, с соответствующими данными для континуума решений Левенберга—Марквардта.

К этому же примеру мы применили алгоритм HFTI. Величины диагональных элементов треугольной матрицы  $R$ , к которой трансформируется матрица  $A$ , равны  $0,52, 0,71 \cdot 10^{-1}, 0,91 \cdot 10^{-2}, 0,14 \cdot 10^{-4}, 0,20 \cdot 10^{-6}$ . Интересно отметить, что эти значения отличаются от соответствующих син-

Таблица 26.2

Нормы решений и невязок метода HFTI для иллюстративной задачи

| $k$ | $\ z^{(k)}\ $ | $\ b - Az^{(k)}\ $ | $k$ | $\ z^{(k)}\ $ | $\ b - Az^{(k)}\ $ |
|-----|---------------|--------------------|-----|---------------|--------------------|
| 1   | 0,99719       | 0,216865           | 4   | 4,92951       | 0,000139           |
| 2   | 2,24495       | 0,039281           | 5   | 220,89008     | 0,000138           |
| 3   | 4,58680       | 0,000139           |     |               |                    |

Таблица 26.3

Нормы решений и невязок при использовании подмножеств столбцов

| Опорные столбцы | $\ w\ $ | $\ b - Aw\ $ | Опорные столбцы | $\ w\ $ | $\ b - Aw\ $ |
|-----------------|---------|--------------|-----------------|---------|--------------|
| 1               | 2,46    | 0,40         | 1, 2, 4         | 10,8    | 0,00018      |
| 2               | 1,92    | 0,22         | 1, 2, 5         | 5,0     | 0,00014      |
| 3               | 2,42    | 0,07         | 1, 3, 4         | 8,1     | 0,00015      |
| 4               | 2,09    | 0,19         | 1, 3, 5         | 5,0     | 0,00014      |
| 5               | 2,30    | 0,19         | 1, 4, 5         | 4,9     | 0,00014      |
| 1, 2            | 5,09    | 0,039        | 2, 3, 4         | 13,5    | 0,00020      |
| 1, 3            | 2,72    | 0,052        | 2, 3, 5         | 7,6     | 0,00014      |
| 1, 4            | 4,53    | 0,023        | 2, 4, 5         | 24,0    | 0,00028      |
| 1, 5            | 5,07    | 0,001        | 3, 4, 5         | 17,3    | 0,00017      |
| 2, 3            | 3,03    | 0,053        | 1, 2, 3, 4      | 10,3    | 0,00014      |
| 2, 4            | 20,27   | 0,030        | 1, 2, 3, 5      | 5,0     | 0,00014      |
| 2, 5            | 17,06   | 0,128        | 1, 2, 4, 5      | 5,0     | 0,00014      |
| 3, 4            | 3,07    | 0,056        | 1, 3, 4, 5      | 5,0     | 0,00014      |
| 3, 5            | 2,97    | 0,058        | 2, 3, 4, 5      | 9,0     | 0,00014      |
| 4, 5            | 17,05   | 0,175        | 1, 2, 3, 4, 5   | 220,9   | 0,00014      |
| 1, 2, 3         | 22,1    | 0,00018      |                 |         |              |

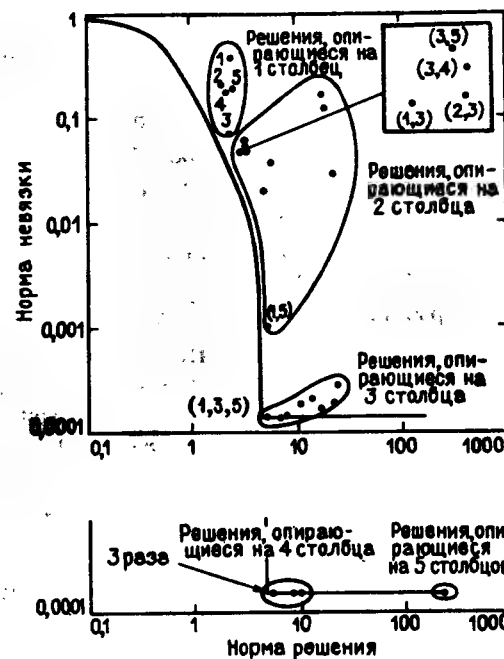


Рис. 26.6. Решения, опирающиеся на подмножества столбцов  $A$

гулярных чисел множителями, не превосходящими 2. Из теоремы 6.31 известно, что сингулярные числа  $s_i$  и диагональные элементы  $r_{ii}$  должны удовлетворять соотношениям

$$1,00 < s_1/r_{11} < 2,24,$$

$$0,58 < s_2/r_{22} < 2,00,$$

$$0,33 < s_3/r_{33} < 1,73,$$

$$0,18 < s_4/r_{44} < 1,41,$$

$$0,09 < s_5/r_{55} < 1,00.$$

Для допуска  $\tau$ , определяющего псевдоранг матрицы, были последовательно установлены значения 0,29, 0,040, 0,0046, 0,0000073, 0,0. Подпрограмма HFTI вычислила пять различных решений (которые мы обозначим через  $z^{(k)}$ ), отвечающих значениям псевдоранга  $k = 1, 2, 3, 4, 5$ . Нормы решений и невязок для этих пяти векторов приведены в табл. 26.2.

Отметим, что данные табл. 26.2 весьма сходны с соответствующими данными рис. 26.1 (в столбцах с шапками "YNORM" и "RNORM") для пробных решений, полученных в сингулярном анализе.

В качестве еще одного способа анализа этой задачи были вычислены решения, опирающиеся на каждое из 31 непустых подмножеств, составленных из пяти столбцов матрицы  $A$ . Нормы решений и невязок для всех этих 31 решений приведены в табл. 26.3 и показаны на рис. 26.6.

Обозначим через  $w^{(i, j, \dots)}$  решение, опирающееся на столбцы  $i, j, \dots$ . Из рис. 26.6 видно, что простейшая форма пошаговой регрессии определит вектор  $w^{(3)}$  как предпочтительное решение среди тех, что опираются только на один столбец. Далее будут рассмотрены решения  $w^{(1,3)}$ ,  $w^{(2,3)}$ ,  $w^{(3,4)}$ ,  $w^{(3,5)}$ . Среди этих четырех векторов будет выбран  $w^{(1,3)}$  как дающий наименьшую норму невязки. Заметим, что эта норма в 52 раза больше, чем та, что соответствует вектору  $w^{(1,5)}$ . Очевидно, что этот последний вектор дает минимум невязки среди всех решений, опирающихся на два столбца. На следующем этапе простая пошаговая регрессия выберет вектор  $w^{(1,3,5)}$ , который как своей нормой, так и нормой соответствующей невязки очень походит на указанные ранее векторы  $x^{(3)}$  и  $z^{(3)}$ .

Дальнейшее поведение пошаговой регрессии критически зависит от деталей конкретного алгоритма. Это связано с очень плохой обусловленностью задач, отвечающих множествам из четырех или пяти столбцов.

## ГЛАВА 27

### МОДИФИКАЦИЯ QR-РАЗЛОЖЕНИЯ ПРИ ДОБАВЛЕНИИ ИЛИ УДАЛЕНИИ СТРОКИ (С ПРИЛОЖЕНИЯМИ К ПОСЛЕДОВАТЕЛЬНОЙ ОБРАБОТКЕ ЗАДАЧ С БОЛЬШИМИ ИЛИ ЛЕНТОЧНЫМИ МАТРИЦАМИ КОЭФФИЦИЕНТОВ)

В этой главе мы адаптируем ортогональные преобразования к последовательной обработке данных для задачи НК. Подобная адаптация является средством экономии машинной памяти для определенного типа задач, связанных с обработкой очень большого объема информации.

Описываемые нами методы имеют важные приложения и к задачам, где нужно получить последовательность решений для массива данных, к которому последовательно добавляется (или из него изымается) информация. Потребность в вычислениях такого рода возникает в классе задач, называемых *последовательным оцениванием* или *фильтрацией*. Методы окаймления, представленные в данной главе, имеют отличную численную устойчивость. Не так обстоит дело во многих опубликованных алгоритмах для этого класса задач, которые основаны на модификации обратной матрицы.

Удаление данных может быть внутренне неустойчивой операцией при определенной связи удаляемых строк со всем информационным массивом. Мы предложим некоторые методы, в которых сделана попытка избежать внесения в задачу дополнительной численной неустойчивости.

Запишем задачу наименьших квадратов обычным образом:

$$Ax \cong b, \quad A = A_m \times n, \quad b = b_m \times 1. \quad (27.1)$$

В § 1 мы рассмотрим задачу, в которой число строк  $A$  велико сравнительно с числом столбцов. В § 2 внимание сосредоточено на случае, когда  $A$

имеет ленточную структуру. Методы, представленные в этих двух параграфах, применимы к задаче последовательного оценивания, так как при необходимости на каждом шаге последовательного накопления легко можно вычислить вектор решения или отвечающую ему ковариационную матрицу. В § 3 демонстрируется применение этих методов на примере выравнивания посредством *линейных сплайнов*. В § 4 описана процедура среднеквадратичной аппроксимации посредством *кубических сплайнов* с равноудаленными узлами. Это еще один пример использования метода ленточного последовательного накопления из § 2.

Методы удаления данных рассматриваются в § 5.

### § 1. Последовательное накопление

Опишем алгоритм преобразования матрицы  $[A : b]$  к верхнему треугольному виду, где не требуется, чтобы вся эта матрица находилась одновременно в памяти машины. Формальное описание этой процедуры дает алгоритм SEQHT (см. 27.10).

Договоримся вначале об обозначениях и приведем неформальное описание процесса. Запишем матрицу  $A$  и вектор  $b$  в блочном виде:

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \dots \\ A_q \end{bmatrix}, \quad (27.2)$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_q \end{bmatrix}. \quad (27.3)$$

Здесь  $A_i$  — матрица размера  $m_i \times n$ , а  $b$  — вектор порядка  $m_i$ . Разумеется,  $m = m_1 + \dots + m_q$ . Числа  $m_i$  могут быть, в частности, равны 1, что позволяет получить большую экономию памяти.

В алгоритме строится последовательность треугольных\*) матриц  $[R_i : d_i]$ ,  $i = 1, \dots, q$ , с тем свойством, что задача наименьших квадратов  $R_i x \cong d_i$  имеет то же множество решений и ту же норму невязки, что и задача

$$\begin{bmatrix} A_1 \\ \dots \\ A_i \end{bmatrix} x \cong \begin{bmatrix} b_1 \\ \dots \\ b_i \end{bmatrix}.$$

Важное обстоятельство, позволяющее экономить память, заключается в том, что матрица  $[R_i : d_i]$  может быть построена и хранима в массиве, где прежде находилась окаймленная матрица

$$[\bar{A}_i : \bar{b}_i] \equiv \begin{bmatrix} R_{i-1} : d_{i-1} \\ A_i : b_i \end{bmatrix}. \quad (27.4)$$

\*) При этом матрицы  $[R_i : d_i]$ , вообще говоря, неквадратные. (Примеч. пер.)

Таким образом, максимальное число хранимых строк равно  $\max_j \{m_j + \min_{i=1}^{j-1} [n+1, \sum_{i=1}^{j-1} m_i]\}$ .

Удобно обозначить через  $[R_0 : d_0]$  (пустую) матрицу, не имеющую строк. В начале  $i$ -го шага алгоритма мы располагаем  $\hat{m}_{i-1} \times (n+1)$ -матрицей  $[R_{i-1} : d_{i-1}]$ , вычисленной на  $i-1$ -м шаге, и новой матрицей данных  $[A_i : b_i]$  размера  $m_i \times (n+1)$ . Положим  $\bar{m}_i = \hat{m}_{i-1} + m_i$ . Составим окаймленную  $\bar{m}_i \times (n+1)$ -матрицу (27.4) и приведем эту матрицу к треугольному виду методом Хаусхолдера:

$$Q_i[\bar{A}_i : \bar{b}_i] = \begin{bmatrix} R_i : d_i \\ 0 : 0 \end{bmatrix} \begin{matrix} \hat{m}_i \\ \bar{m}_i - \hat{m}_i \end{matrix} \quad (27.5)$$

Здесь  $\hat{m}_i = \min\{n+1, \bar{m}_i\}$ .

Теперь  $i$ -й шаг алгоритма завершен. Для последующих ссылок обозначим результат  $q$  шагов алгоритма через

$$\begin{bmatrix} \hat{R} & \hat{d} \\ 0 & e \end{bmatrix} \begin{matrix} n \\ 1 \end{matrix} \equiv [R_q : d_q]. \quad (27.6)$$

Легко видеть, что существует ортогональная матрица  $Q$  такая, что

$$Q[A : b] = \begin{bmatrix} R & d \\ 0 & e \\ 0 & 0 \end{bmatrix}.$$

Следовательно, задача наименьших квадратов

$$\begin{bmatrix} R \\ 0 \end{bmatrix} x \cong \begin{bmatrix} d \\ e \end{bmatrix} \quad (27.7)$$

эквивалентна исходной задаче (27.1) в том смысле, что обе имеют одинаковое множество решений и один и тот же минимум нормы невязки. Кроме того, матрица  $R$  имеет те же сингулярные числа, что и матрица  $A$  из (27.1).

Опишем теперь вычислительный алгоритм, который формализует эту процедуру. С этой целью положим

$$\nu_j = \begin{cases} 0, & j=0, \\ \sum_{i=1}^j m_i, & j>0, \end{cases} \quad (27.8)$$

$$\mu = \max_{1 \leq j \leq q} \{m_j + \min[n+1, \nu_{j-1}]\}. \quad (27.9)$$

Весь процесс обработки будет происходить в массиве  $W$  машинной памяти, состоящем по крайней мере из  $\mu$  строк и  $n+1$  столбцов. Обозначение  $W(i_1 : i_2, j_1 : j_2)$  будет относиться к подмассиву массива  $W$ , образованному пересечением строк  $i_1, \dots, i_2$  и столбцов  $j_1, \dots, j_2$ . Через  $W(i, j)$  будем обозначать элемент  $(i, j)$  массива  $W$ . Символ  $\rho$  указывает отдельную ячейку памяти.

**А л г о р и т м 27.10. SEQHT (последовательная триангуляризация Хаусхолдера):**

1. Положить  $l := 0$ .
2. Для  $t := 1, \dots, q$  выполнить шаги 3–6.
3. Положить  $p := l + m_t$ .
4. Положить  $W(l+1 : p, 1 : n+1) := [A_t : b_t]$  (см. (27.2) и (27.3)).
5. Для  $i := 1, \dots, \min(n+1, p-1)$  выполнить алгоритм H1( $i, \max(i+1, l+1), p, W(1, i), \rho, W(1, i+1), n-i+1$ ).
6. Положить  $l := \min(n+1, p)$ .
7. **Замечание.** Матрица  $[A : b]$  приведена к верхней треугольной форме (27.6).

Отметим, что для шага 4 нужно одновременно вводить в память машины лишь подматрицу  $[A_t : b_t]$ . Вся информация может быть обработана в рабочем массиве  $W$  размера  $\mu \times (n+1)$ . Усложняя программирование, можно еще больше сократить запросы к памяти, используя то обстоятельство, что матрица  $[R_j : d_j]$  верхняя треугольная.

Для последующего обсуждения числа операций обозначим через  $\alpha$  сложение или вычитание, а через  $\mu$  умножение или деление. Положим

$$\nu = \frac{mn^2 - n^3/3}{2}.$$

В табл. 19.1 было указано, что число операций при триангуляризации  $m \times n$ -матрицы ( $m > n$ ) посредством преобразований Хаусхолдера равно приблизительно  $\nu(2\alpha + 2\mu)$ , если игнорировать при подсчете константы и члены, линейные или квадратичные по  $m$  и  $n$ . Если алгоритм Хаусхолдера применяется последовательно в  $q$  этапов, как в алгоритме SEQHT (см. 27.10), то число операций возрастает примерно до  $\nu(2\alpha + 2\mu)(m+q)/m$ . Если вводимые блоки данных состоят каждый из  $k$  строк ( $kq = m$ ), то число операций можно записать формулой  $\nu(2\alpha + 2\mu)(k+1)/k$ .

Таким образом, стоимость последовательного накопления в методе Хаусхолдера возрастает по мере уменьшения размера блока. В предельном случае  $k = 1$  число операций приблизительно удваивается по сравнению с обычным алгоритмом Хаусхолдера.

При малых размерах блоков может оказаться более экономичной замена преобразований Хаусхолдера на шаге 5 алгоритма SEQHT (см. 27.10) каким-либо методом, основанным на двумерных вращениях или отражениях. Число операций в этих последних методах не зависит от размера блока. Если для триангуляризации используется метод Гивенса (алгоритмы G1 (см. 10.25) и G2 (см. 10.26)), то число операций составит  $\nu(2\alpha + 4\mu)$ . Для хаусхолдеровых  $2 \times 2$ -преобразований, записанных в виде (10.27), число операций равно  $\nu(3\alpha + 3\mu)$ .

Если метод Гивенса применяется в модификации Джентльмена (см. гл. 10), то число операций снижается до  $\nu(2\alpha + 2\mu)$ . Следовательно, этот метод может составить конкуренцию обычному (непоследовательному) методу Хаусхолдера и требует меньше арифметических операций, чем модификация метода Хаусхолдера для последовательной обработки.

На практике соотношение характеристик программ, реализующих эти методы, сильно зависит от деталей программирования.

после того как алгоритм SEQHT (см. 27.10) закончил работу, можно вычислить (если матрица  $R$  в (27.6) не вырождена) решение  $\hat{x}$  из системы

$$Rx = d. \quad (27.11)$$

Для числа  $e$  из (27.6)

$$|e| = \|\hat{Ax} - b\|. \quad (27.12)$$

Во многих приложениях нужна нешкалированная матрица ковариации (см. (12.1))

$$C = (A^T A)^{-1} = R^{-1} (R^{-1})^T. \quad (27.13)$$

Вычисление  $R^{-1}$  можно осуществить на том месте, которое занимает  $R$ , а затем на том же месте можно вычислить (верхнюю треугольную часть)  $R^{-1} (R^{-1})^T$ . Тем самым матрица  $C$  из (27.13) может быть вычислена, по существу, без использования дополнительной памяти. Более подробно организация этих вычислений обсуждается в гл. 12. Кроме того можно вычислить величину (см. (12.2))

$$\sigma^2 = \frac{e^2}{m - n},$$

где число  $e$  определено в (27.6).

Для определенности алгоритм SEQHT был оформлен в виде цикла, выполняемого при  $t = 1, \dots, q$ . В реальных приложениях этой последовательной обработки более вероятно, что шаги 3–6 алгоритма будут реализованы как подпрограмма. Число  $q$  и весь массив данных, образующий матрицу  $[A : b]$ , могут быть поначалу неизвестны. Чтобы найти решение, опирающееся на накопленную до сих пор информацию, вызывающая программа может использовать текущую треугольную матрицу  $R$  на любом этапе, где  $R$  не вырождена. Если имеется дополнительный массив в  $n(n+1)/2$  машинных слов, то вызывающая программа может вычислить и верхнюю треугольную часть нешкалированной матрицы ковариации  $R_t^{-1} (R_t^{-1})^T$ , опять же в предположении, что  $R_t$  не вырождена. Таким образом, шаги 3–6 алгоритма SEQHT составляют его ядро, на основе которого можно строить программы последовательного оценивания.

## § 2. Последовательное накапливание ленточных матриц

В некоторых задачах матрица исходных данных  $[A : b]$  имеет или после предварительной перестановки строк и столбцов приобретает следующую ленточную структуру. Существуют целое число  $n_b$ ,  $n_b \leq n$ , и неубывающая последовательность целых чисел  $j_1, \dots, j_q$ , такие, что все ненулевые элементы подматрицы  $A_t$  сосредоточены в столбцах  $j_t, \dots, j_t + n_b - 1$ . Таким образом,  $A_t$  имеет вид

$$A_t = \left[ \begin{array}{c|c|c} 0 & : C_t & 0 \\ \hline j_t - 1 & n_b & n - n_b - j_t + 1 \end{array} \right] m_t. \quad (27.14)$$

Число  $n_b$  мы будем называть шириной ленты матрицы  $A$ .

Легко проверить, что все ненулевые элементы  $i$ -й строки верхней треугольной матрицы  $R$  из (27.6) находятся в столбцах  $i, \dots, i + n_b - 1$ .

Кроме того, строки  $R$  с номерами  $1, \dots, j_t - 1$  остаются неизменными, когда алгоритм 27.10 обрабатывает подматрицы  $A_1, \dots, A_q$ .

Эти замечания позволяют модифицировать алгоритм 27.10 таким образом, что достаточно будет рабочего массива с  $n_b + 1$  столбцами. Итак, пусть  $G$  – массив размера  $\mu \times (n_b + 1)$ , где  $\mu$  удовлетворяет соотношению (27.9). В алгоритме рабочий массив  $G$  разбивается на три подмассива  $G_1, G_2$  и  $G_3$ ; именно:

$$G_1 = \text{строки } G \text{ с номерами } 1, \dots, i_p - 1; \quad (27.15)$$

$$G_2 = \text{строки } G \text{ с номерами } i_p, \dots, i_r - 1; \quad (27.16)$$

$$G_3 = \text{строки } G \text{ с номерами } i_r, \dots, i_r + m_t - 1. \quad (27.17)$$

Целые числа  $i_p$  и  $i_r$  определяются алгоритмом, и их значения изменяются в ходе обработки. Границы их изменения устанавливаются неравенствами  $1 \leq i_p \leq i_r \leq n + 2$ .

На разных этапах алгоритма столбец  $G$  с номером  $n_b + 1$  содержит либо вектор  $\bar{b}_t$  (см. правую часть (27.4)), либо обработанный вектор  $d_t$  (см. правую часть (27.5)).

Для  $1 \leq j \leq n$  соответствие между элементами массива и матричными элементами таково:

$$\text{в } G_1: \text{ элемент } (i, j) \text{ массива содержит матричный элемент } (i, i + j - 1); \quad (27.18)$$

$$\text{в } G_2: \text{ элемент } (i, j) \text{ массива содержит матричный элемент } (i, i_p + j - 1); \quad (27.19)$$

$$\text{в } G_3: \text{ элемент } (i, j) \text{ массива содержит матричный элемент } (i, j_t + j - 1). \quad (27.20)$$

Рисунки 27.1 и 27.2 поясняют основную идею алгоритма компактного хранения для ленточной задачи наименьших квадратов. Если бы алгоритм 27.10 был применен к блочно-диагональной задаче с  $n_b = 4$ , то на том шаге, где в рабочий массив вводится блок данных  $[C_t : b_t]$ , ситуация сходна с показанной на рис. 27.1. Используя ленточную структуру, можно упаковать эту информацию в соответствии с рис. 27.2.

В момент, показанный на рисунках, выполняются неравенства  $i_p \leq j_t \leq i_r$ . Дополнительные графические иллюстрации к этому алгоритму можно найти в § 3 (см. рис. 27.4 и 27.5).

Теперь приведем подробное описание алгоритма.

Будем использовать такие обозначения:  $G(i, j)$  – элемент  $(i, j)$  массива  $G$ ,  $G(i_1 : i_2, j_1 : j_2)$  – подмассив  $G$ , образованный элементами  $G(i, j)$  с индексами  $i_1 \leq i \leq i_2, j_1 \leq j \leq j_2$ .

Алгоритм 27.21. BSEQHT (ленточная последовательная триангуляризация Хаусхолдера):

1. Положить  $i_r := 1, i_p := 1$ .

2. Для  $t := 1, \dots, q$  выполнить шаги 3–24.

3. Замечания. В этот момент нужно передать алгоритму данные  $[C_t : b_t]$  и целые числа  $m_t$  и  $j_t$ . Предполагается, что  $m_t > 0$  при всех  $t$  и  $j_q > j_{q-1} > \dots > j_1 > 1$ .

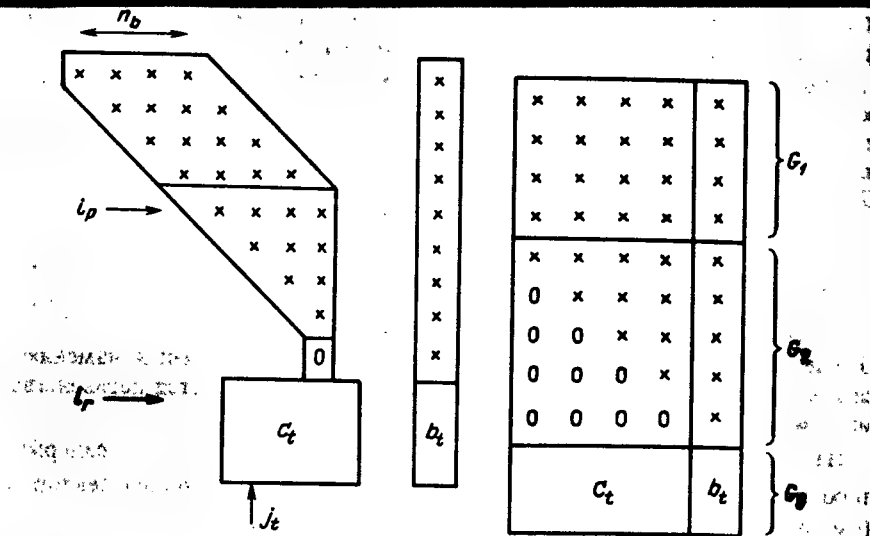


Рис. 27.1. Введение блока данных  $[C_t; b_t]$

Рис. 27.2. Те же данные, что и на рис. 27.1, упакованы в массиве  $G$ .

4. Положить  $G(i_r : i_r + m_t - 1, 1 : n_b + 1) := [C_t; b_t]$ . (Определение  $C_t$  дано в (27.14). Заметим, что данные вводятся в подмассив  $G_3$  массива  $G$ .)
5. Если  $j_t = i_p$ , перейти к шагу 18.
6. *Замечание.* Предположение о монотонности чисел  $j_t$  гарантирует, что здесь  $j_t > i_p$ .
7. Если  $j_t \leq i_r$ , перейти к шагу 12.
8. *Замечание.* Здесь  $j_t$  больше, чем  $i_r$ . Это свидетельствует о неполноте ранга, так как диагональный элемент  $(i_r, i_r)$  треугольной матрицы  $R$  из (27.6) будет нулем. Тем не менее триангуляризацию можно завершить. Некоторые методы решения этой задачи наименьших квадратов с неполным рангом будут рассмотрены вслед за описанием алгоритма.
9. Переместить содержимое  $G(i_r : i_r + m_t - 1, 1 : n_b + 1)$  в  $G(j_t : j_t + m_t - 1, 1 : n_b + 1)$ .
10. Присвоить нулевые значения всем элементам подмассива  $G(i_r : j_t - 1, 1 : n_b + 1)$ .
11. Положить  $i_r := j_t$ .
12. Положить  $\mu := \min(n_b - 1, i_r - i_p - 1)$ ; если  $\mu = 0$ , перейти к шагу 17.
13. Для  $l := 1, \dots, \mu$  выполнить шаги 14–16.
14. Положить  $k := \min(l, j_t - i_p)$ .
15. Для  $i := l + 1, \dots, n_b$  положить  $G(i_p + l, i - k) := G(i_p + l, i)$ .
16. Для  $i := 1, \dots, k$  положить  $G(i_p + l, n_b + 1 - i) := 0$ .
17. Положить  $i_p := j_t$ . (Заметим, что это присвоение переопределяет границу между подмассивами  $G_1$  и  $G_2$  из (27.15) и (27.16).)
18. *Замечание.* Шаги 19, 20 — это применение хаусхолдовой триангуляризации к строкам с номерами  $i_p, \dots, i_r + m_t - 1$ . Как отмечалось в об-

суждении алгоритма SEQHT (см. 27.10), некоторого сокращения времени счета можно добиться, заменяя общий метод Хаусхолдера методами, использующими двумерные вращения или отражения.

19. Положить  $\hat{m} := i_r + m_t - i_p$ . Положить  $\hat{k} := \min(n_b + 1, \hat{m})$ .
20. Для  $i := 1, \dots, \hat{k}$  выполнить алгоритм H1( $i, \max(i + 1, i_r - i_p + 1), \hat{m}, G(i_p, i), \rho, G(i_p, i + 1), n_b + 1 - i$ ).
21. Положить  $i_r := i_p + \hat{k}$ . (Заметим, что это присвоение переопределяет границу между подмассивами  $G_2$  и  $G_3$  из (27.16) и (27.17).)
22. Если  $\hat{k} < n_b + 1$ , перейти к шагу 24.
23. Для  $j := 1, \dots, n_b$  положить  $G(i_r - 1, j) := 0$ .
24. Вернуться на начало цикла.
25. *Замечание.* Основной цикл закончен. Треугольная матрица из (27.6) хранится в подмассивах  $G_1$  и  $G_2$  (см. (27.15) и (27.16)) в соответствии с правилами (27.18) и (27.19).  
Если все диагональные элементы матрицы  $R$  из (27.6) ненулевые, то решение задачи (27.7) можно найти обратной подстановкой, описываемой шагами 26–31. Заметим, что диагональные элементы  $R$  используются как делители на шаге 31. Обсуждение некоторых методов для вырожденного случая приводится вслед за описанием алгоритма.

26. Для  $i := 1, \dots, n$  положить  $X(i) := G(i, n_b + 1)$ .
27. Для  $i := n, n - 1, \dots, 1$  выполнить шаги 28–31.
28. Положить  $s := 0$  и  $l := \max(0, i - i_p)$ .
29. Если  $i = n$ , перейти к шагу 31.
30. Для  $j := 2, \dots, \min(n + 1 - i, n_b)$  положить  $s := s + G(i, j + 1) \times X(i - 1 + j)$ .
31. Положить  $X(i) := [X(i) - s] / G(i, l + 1)$ .
32. *Замечание.* В массиве  $X$  теперь хранится решение  $x$ . Если (как обычно и бывает) полная информационная матрица  $[A; b]$  (см. (27.2) и (27.3)) имеет более чем  $n$  строк, то скаляр  $e$  (см. (27.6)) будет храниться в ячейке  $G(n + 1, n_b + 1)$ . Модуль  $e$  равен норме невязки, отвечающей решению  $x$ .  
Ленточная структура  $A$  в общем случае не означает, что нешкалированная матрица ковариации  $C$  из (12.1) также будет иметь какую-либо ленточную структуру. Однако столбцы  $c_j, j = 1, \dots, n$ , этой матрицы можно вычислить по одному, не используя дополнительной памяти. Именно, вектор  $c_j$  можно найти, решая две ленточные треугольные системы

$$R^T w_j = e_j, \quad (27.22)$$

$$R c_j = w_j, \quad (27.23)$$

где  $e_j$  —  $j$ -й столбец единичной матрицы порядка  $n$ .

Алгоритм 27.24 описывает шаги, необходимые для решения системы (27.22) или, более общо, системы  $R^T u = z$ , где  $z$  — произвольный вектор. Предполагается, что матрица  $R$  хранится в массиве  $G$  так же, как на выходе алгоритма 27.21. Вектор правой части, например  $e_j$  в (27.22), должен быть помещен в массив  $X$ . В результате выполнения алгоритма 27.24 исходное содержимое массива  $X$  будет заменено решением системы. В случае (27.22) это будет вектор  $w_j$ .



Алгоритм 27.24. Решение системы  $R^T y = z$ :

1. Для  $j := 1, \dots, n$  выполнить шаги 2–6.
2. Положить  $s := 0$ .
3. Если  $j = 1$ , перейти к шагу 6.
4. Положить  $i_1 := \max(1, j - n_b + 1)$ . Положить  $i_2 := j - 1$ .
5. Для  $i := i_1, \dots, i_2$  положить  $s := s + X(i) \times G(i, j - i + 1 + \max(0, i - i_p))$ .
6. Положить  $X(j) := (X(j) - s) / G(j, 1 + \max(0, j - i_p))$ .

После того как из (27.22) определен вектор  $w_j$ , решить систему (27.23) относительно  $c_j$  можно, выполняя шаги 27–31 алгоритма 27.21.

Далее можно найти величину (см. (12.2))

$$\sigma^2 = \frac{e^2}{m - n},$$

где  $e_j$  определено в (27.6). Величина  $e$  вычисляется алгоритмом BSEQHT, как указано в замечании на шаге 32.

Замечания в конце § 1, относящиеся к возможным реализациям алгоритма SEQHT (см. 27.10) в случае последовательного оценивания, применимы с очевидными изменениями также и к алгоритму BSEQHT. Есть, однако, и существенное различие: в алгоритме BSEQHT число столбцов, имеющих ненулевые элементы, обычно возрастает по мере ввода новых блоков данных. Поэтому в типичном случае на ранних этапах процесса можно определить меньшее число компонент решения, чем на последующих.

На шагах 8 и 25 упоминалось о том, что задача может иметь неполный ранг. Наш опыт показывает, что в этой ситуации весьма хорошо работает техника Левенберга (гл. 25, § 4), причем она не увеличивает ширину ленты задачи. Мы используем ниже обозначения формул (25.30)–(25.34).

Пусть матрица  $\tilde{A}$  в (25.31) ленточная с шириной ленты  $n_b$ . Сейчас мы предположим, что и  $F$  ленточная и ее ширина ленты не превосходит  $n_b$ . На практике в качестве  $F$  обычно берут диагональную матрицу. Подходящими перестановками строк матрица

$$\begin{bmatrix} \tilde{A} & \tilde{b} \\ \lambda F & \lambda d \end{bmatrix}$$

из (25.31) может быть преобразована в новую матрицу  $[\hat{A} : \hat{b}]$ , где  $\hat{A}$  имеет ширину ленты  $n_b$ . К матрице  $[\hat{A} : \hat{b}]$  можно затем применить алгоритм BSEQHT. Если  $F$  не вырождена, то  $\hat{A}$  имеет полный ранг. Если к тому же  $\lambda$  достаточно велико, то у  $\hat{A}$  будет и полный псевдоранг. Поэтому алгоритм BSEQHT можно завершить, включая вычисление решения.

Если нужно исследовать зависимость решения задачи (25.31) от  $\lambda$ , то можно вначале обработать только подматрицу  $[\tilde{A} : \tilde{b}]$ , приведя ее к ленточному треугольному виду

$$T = \begin{bmatrix} R : \tilde{d} \\ 0 : e \end{bmatrix}.$$

Будем считать, что эта матрица  $T$  хранится в памяти машины (если необходимо, то и во внешней).

Чтобы решить задачу (25.31) для конкретного значения  $\lambda$ , заметим, что эквивалентная задача имеет вид

$$\begin{bmatrix} R \\ 0 \\ \lambda F \end{bmatrix} x \cong \begin{bmatrix} \tilde{d} \\ e \\ \lambda d \end{bmatrix}.$$

Теперь можно перестановками строк добиться, чтобы матрица коэффициентов имела ширину ленты  $n_b$ . Решение этой преобразованной задачи можно вычислить по алгоритму BSEQHT.

Этот прием перемешивания строк матрицы  $[\lambda F : \lambda d]$  со строками предварительно вычисленной треугольной матрицы  $[R : \tilde{d}]$ , для того чтобы сохранить ширину ленты, можно использовать также и как метод введения новых уравнений в ранее решенную задачу. Он позволяет при решении расширенной задачи избежать повторной триангуляризации исходного массива данных.

В случае, когда  $F = I_n$  и  $d = 0$ , выбор параметра  $\lambda$  можно упростить, если вычислить сингулярные числа и преобразованную правую часть задачи наименьших квадратов (27.7). Именно, если  $R = USV^T$  — сингулярное разложение матрицы  $R$ , то вектор

$$g = U^T \tilde{d} \quad (27.25)$$

и матрицу  $S = \text{diag} \{s_1, \dots, s_n\}$  можно вычислить, не используя никаких других массивов, кроме того, где находится ленточная матрица  $R$ . Основные этапы этого вычисления таковы. Матрица  $R$  умножается слева и справа на конечные последовательности вращений Гивенса  $J_i$  и  $T_i$ ; в результате этих умножений матрица  $B = T_n \dots T_1 R J_1 \dots J_n$  становится двухдиагональной. Произведение  $T_n \dots T_1 \tilde{d}$  замещает в памяти вектор  $\tilde{d}$ . Посредством алгоритма QRBD (см. 18.31) вычисляется сингулярное разложение матрицы  $B$ . Соответствующие левые вращения применяются и к прежнему вектору  $\tilde{d}$ ; в конечном счете будет получен вектор  $g$  из (27.25). Значение  $\lambda$  можно теперь определить из условия, чтобы норма невязки или норма решения были равны заданным числам; при этом используются формулы (25.48) или (25.47) соответственно.

### § 3. Пример: линейные сплайны

Чтобы закрепить идеи, изложенные в §§ 1, 2, рассмотрим следующую задачу об аппроксимации данных (об информационном сжатии). Пусть заданы  $m$  пар  $\{(t_i, y_i)\}$ , абсциссы  $t_i$ ,  $i = 1, \dots, m$ , которых находятся на отрезке  $[a, b]$ . Нужно аппроксимировать эти данные функцией  $f(t)$ , представление которой требует меньшей памяти, чем сами данные. Простейшей непрерывной функцией (обладающей некоторой степенью общности), которую можно использовать для аппроксимации в смысле наименьших квадратов, является, по-видимому, кусочно линейная непрерывная функция, определяемая следующим образом.

Отрезок  $[a, b]$  разбивается на  $n - 1$  отрезков узлами  $t^{(i)}$ , так что

$$a = t^{(1)} < t^{(2)} < \dots < t^{(n)} = b. \quad (27.26)$$

Для значений  $t$  на отрезке

$$t^{(i)} \leq t \leq t^{(i+1)} \quad (27.27)$$

положим

$$w_i(t) = \frac{t - t^{(i)}}{t^{(i+1)} - t^{(i)}}, \quad (27.28)$$

$$z_i(t) = 1 - w_i(t) = \frac{t^{(i+1)} - t}{t^{(i+1)} - t^{(i)}}, \quad (27.29)$$

$$f(t) = x_i z_i(t) + x_{i+1} w_i(t), \quad i = 1, \dots, n-1. \quad (27.30)$$

Параметры  $x_i, i = 1, \dots, n$ , формул (27.30) должны быть определены как переменные линейной задачи наименьших квадратов.

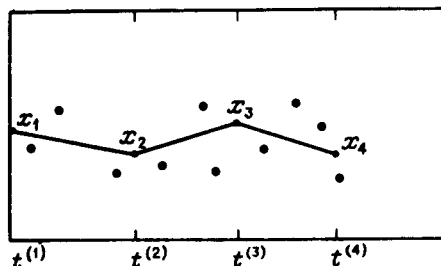


Рис. 27.3. Пример линейного сплайна для  $m = 10, n = 4$ . Точками указаны заданные пары  $(t_i, y_i)$

Для примера возьмем  $m = 10$  и  $n = 4$ , причем узлы будем считать равноудаленными. Схематический график такой задачи представлен на рис. 27.3.

Для коэффициентов  $x_i$  получается задача наименьших квадратов, форма которой показана на рис. 27.4. Отметим, что матрица этой задачи ленточная с шириной ленты  $n_b = 2$ .

Рис. 27.5 демонстрирует работу ленточного алгоритма BSEQHT (см. 27.21) для этого примера.

(1) Вначале  $i_p = i_r = 1$ . Вводится первый блок данных, образованный нетривиальными элементами первых трех строк на рис. 27.4. Полагаем  $j_1 = 1, m_1 = 3$ .

(2) Первый блок приводится к треугольному виду посредством преобразований Хаусхолдера. Полагаем  $i_r = 4$ .

(3) Вводим второй блок данных рис. 27.4. Полагаем  $j_2 = 2, m_2 = 3$ .

(4) Левый сдвиг второй строки без ее последнего элемента, который относится к вектору правой части. Полагаем  $i_p = j_2 (= 2)$ .

(5) Триангуляризация второй — шестой строк. Полагаем  $i_r = 5$ .

(6) Вводим третий блок данных рис. 27.4. Полагаем  $j_3 = 3, m_3 = 4$ .

(7) Левый сдвиг третьей строки без ее последнего элемента. Полагаем  $i_p = j_3 (= 3)$ .

(8) Триангуляризация строк третьей — восьмой. Полагаем  $i_r = 6$ .

Результатам, полученным на этапе (8), соответствует задача наименьших квадратов, показанная на диаграмме (9). Эту задачу можно теперь решить обратной подстановкой.

$$\begin{bmatrix} z_1(t_1) & w_1(t_1) & & & \\ z_1(t_2) & w_1(t_2) & & & \\ z_1(t_3) & w_1(t_3) & & & \\ 0 & z_2(t_4) & w_2(t_4) & & \\ 0 & z_2(t_5) & w_2(t_5) & & \\ 0 & z_2(t_6) & w_2(t_6) & & \\ 0 & 0 & z_3(t_7) & w_3(t_7) & \\ 0 & 0 & z_3(t_8) & w_3(t_8) & \\ 0 & 0 & z_3(t_9) & w_3(t_9) & \\ 0 & 0 & z_3(t_{10}) & w_3(t_{10}) & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix}$$

Рис. 27.4

$$\begin{array}{lll} (1) \begin{bmatrix} a & a & a \\ a & a & a \\ a & a & a \end{bmatrix} & \rightarrow & (2) \begin{bmatrix} b & b & b \\ 0 & b & b \\ 0 & 0 & b \end{bmatrix} \\ (3) \begin{bmatrix} b & b & b \\ 0 & b & b \\ 0 & 0 & b \\ c & c & c \\ c & c & c \\ c & c & c \end{bmatrix} & \rightarrow & (4) \begin{bmatrix} b & b & b \\ b & 0 & b \\ 0 & 0 & b \\ c & c & c \\ c & c & c \\ c & c & c \end{bmatrix} & (5) \begin{bmatrix} b & b & b \\ d & d & d \\ 0 & d & d \\ 0 & 0 & d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ (6) \begin{bmatrix} b & b & b \\ d & d & d \\ 0 & d & d \\ 0 & 0 & d \\ e & e & e \\ e & e & e \\ e & e & e \end{bmatrix} & \rightarrow & (7) \begin{bmatrix} b & b & b \\ d & d & d \\ d & 0 & d \\ 0 & 0 & d \\ e & e & e \\ e & e & e \\ e & e & e \end{bmatrix} & (8) \begin{bmatrix} b & b & b \\ d & d & d \\ f & f & f \\ 0 & f & f \\ 0 & 0 & f \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ (9) \begin{bmatrix} b & b & 0 & 0 \\ 0 & d & d & 0 \\ 0 & 0 & f & f \\ 0 & 0 & 0 & f \\ 0 & 0 & 0 & 0 \end{bmatrix} & \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = & \begin{bmatrix} b \\ d \\ f \\ f \end{bmatrix} \end{array}$$

Рис. 27.5

Иллюстрацией размера экономии памяти, которую дает такая ленточная обработка, служит пример аппроксимации линейным сплайном для  $m = 1000$  точек и 100 отрезков. Таким образом, задача наименьших квадратов имеет  $n = 101$  неизвестных. Предположим еще, что строчная размерность каждого обрабатываемого блока не превышает 10. Тогда максимальные размеры рабочего блока не превосходят  $[n + 1 + \max(m_i)] \times 3 = (101 + 1 + 10) \times 3$ , т.е. 336 ячеек. Менее специализированный алгоритм последовательного накопления (типа алгоритма из § 1) потребовал бы рабочего массива с размерами по крайней мере  $[n + 1 + \max(m_i)] \times (n + 1) = (101 + 1 + 10) \times 102$ , т.е. 11 424 ячеек. Если бы все строки были введены одновременно то рабочий массив имел бы размер  $m \times (n + 1) = 1000 \times 101$  и состоял из 101 000 ячеек.

#### § 4. Сглаживание посредством кубических сплайнов

В качестве еще одного примера использования последовательной обработки в ленточном случае мы обсудим задачу о сглаживании данных посредством кубических сплайнов с равноудаленными узлами.

Пусть заданы числа  $b_1 < b_2 < \dots < b_n$ . Через  $S$  обозначим множество всех кубических сплайнов, определенных на отрезке  $[b_1, b_n]$  и имеющих внутренние узлы  $b_2, \dots, b_{n-1}$ . Функция  $f$  тогда и только тогда принадлежит  $S$ , когда она является кубическим многочленом на каждом отрезке  $[b_k, b_{k+1}]$ , а на всем отрезке  $[b_1, b_n]$  непрерывна вместе со своей первой и второй производными.

Можно показать, что множество  $S$  является  $n + 2$ -мерным линейным пространством. Поэтому любая система  $\{q_j : j = 1, \dots, n + 2\}$  линейно независимых функций из  $S$  будет в  $S$  базисом. Это значит, что каждая функция  $f \in S$  имеет единственное представление вида  $f(x) = \sum_{j=1}^{n+2} c_j q_j(x)$ .

Задача об отыскании элемента  $S$ , который наилучшим образом (в смысле метода наименьших квадратов) аппроксимирует заданную информацию  $\{(x_i, y_i) : x_i \in [b_1, b_n]; i = 1, \dots, m\}$ , принимает с помощью базиса  $\{q_j\}$  форму  $Ac \approx y$ , где  $a_{ij} = q_j(x_i)$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n + 2$ . Здесь  $c$  —  $n + 2$ -вектор с компонентами  $c_j$ , а  $y$  —  $m$ -вектор с компонентами  $y_i$ .

Пусть заданные точки упорядочены так, что  $x_1 \leq \dots \leq x_m$ . В  $S$  существуют базисы с тем свойством, что соответствующие им матрицы  $A$  ленточные с шириной ленты 4. Методы вычисления таких базисных функций для равноудаленных узлов описаны в [36, 41, 45, 46]. Чтобы избежать усложнений, не существенных для иллюстрации ленточной последовательной

Таблица 27.1

Исходные данные для сглаживания посредством кубических сплайнов

| $x$ | $y$ | $x$ | $y$ |
|-----|-----|-----|-----|
| 2   | 2,2 | 14  | 3,8 |
| 4   | 4,0 | 16  | 5,1 |
| 6   | 5,0 | 18  | 6,1 |
| 8   | 4,6 | 20  | 6,3 |
| 10  | 2,8 | 22  | 5,0 |
| 12  | 2,7 | 24  | 2,0 |

Таблица 27.2

RMS как функция от NBP

| NBP | 5     | 6     | 7     | 8     | 9     | 10  |
|-----|-------|-------|-------|-------|-------|-----|
| RMS | 0,254 | 0,085 | 0,134 | 0,091 | 0,007 | 0,0 |

обработки, мы ограничимся здесь обсуждением только случая равноудаленных узлов.

Чтобы построить нужный базис, положим  $h = b_{k+1} - b_k$ ,  $k = 1, \dots, n - 1$ . Введем два кубических многочлена:

$$p_1(t) = 0,25t^3, \quad p_2(t) = 1 - 0,75(1+t)(1-t)^2.$$

Пусть  $I_1$  обозначает замкнутый интервал  $[b_1, b_2]$ , а  $I_k$  — полуоткрытый интервал  $(b_k, b_{k+1}]$ ,  $k = 2, \dots, n - 1$ .

В интервале  $I_k$  только четыре из функций  $q_j$  принимают ненулевые значения. Эти четыре функции определяются для  $x \in I_k$  формулами

$$t = \frac{x - b_k}{h},$$

$$q_k(x) = p_1(1-t), \quad q_{k+1}(x) = p_2(1-t),$$

$$q_{k+2}(x) = p_2(t), \quad q_{k+3}(x) = p_1(t).$$

Рассмотрим иллюстративную задачу сглаживания, используя эту систему базисных функций и применяя к полученной ленточной матрице последо-

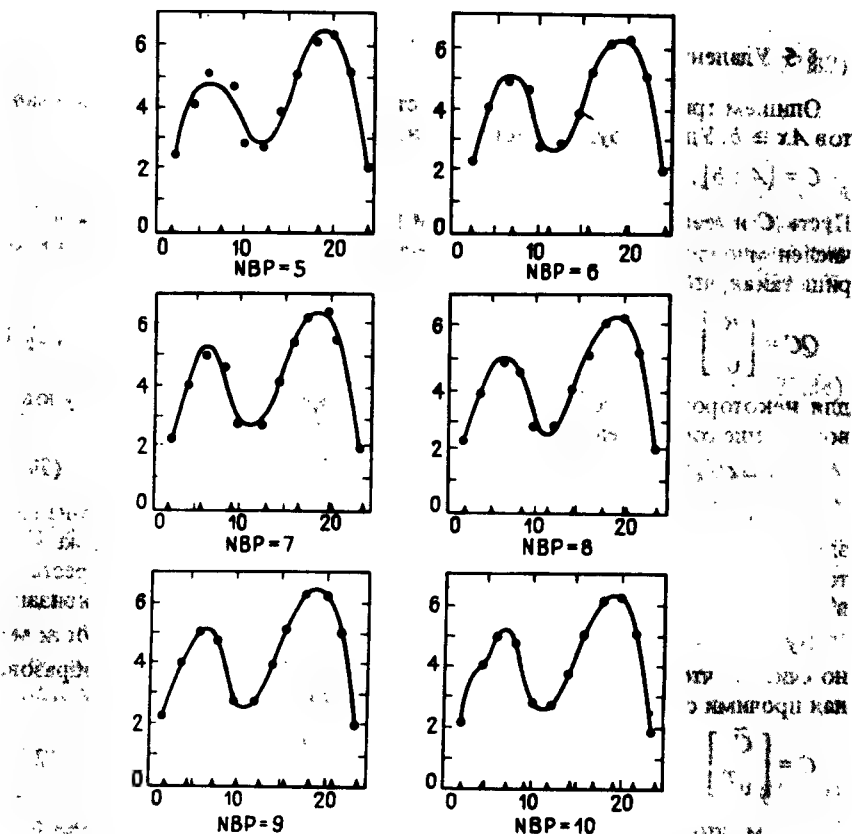


Рис. 27.4. Сглаживающие сплайны для данных табл. 27.1. Треугольники у основания каждого графика указывают абсциссы узлов

вательное хаусхолдерово накопление. Исходные данные для этого примера приведены в табл. 27.1.

Число узлов указывает параметр NBP. Ему последовательно были присвоены значения 5, 6, 7, 8, 9, 10. При фиксированном значении NBP абсциссы узлов определяются формулами

$$b_i = 2 + \frac{22(i-1)}{NBP-1}, \quad i = 1, \dots, NBP.$$

Количество вычисляемых коэффициентов равно  $NC = NBP + 2$ . Заметим, что в случае  $NBP = 6$  число коэффициентов совпадает с числом заданных точек. В этом случае сглаживающая кривая интерполирует эти точки. Положим

$$RMS = \left( \frac{1}{12} \sum_{i=1}^{12} r_i^2 \right)^{1/2},$$

где  $r_i$  — невязка в  $i$ -й точке. Значение RMS для каждого случая указано в табл. 27.2. Отметим, что RMS не является монотонной функцией от NBP.

График каждой из шести сглаживающих кривых приведен на рис. 27.6.

## § 5. Удаление строк

Опишем три метода для удаления строки из задачи наименьших квадратов  $Ax \cong b$ . Удобно будет ввести обозначение

$$C = [A : b]. \quad (27.31)$$

Пусть  $C$  имеет  $m$  строк,  $n$  столбцов и ранг  $k$ . Предположим, что для  $C$  вычислен множитель Холесского  $R$ . Здесь  $R$  — верхняя треугольная  $k \times n$ -матрица такая, что

$$QC = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad (27.32)$$

для некоторой ортогональной матрицы  $Q$  порядка  $m$ . Матрица  $R$  удовлетворяет еще соотношению

$$R^T R = C^T C. \quad (27.33)$$

Удобно сделать дополнительное предположение, что все  $k$  диагональных элементов  $R$  ненулевые. Так обязательно будет, если  $k = n$ . Если же  $k < n$ , то невырожденности диагональных элементов можно добиться, переставляя в случае необходимости столбцы и проводя новую триангуляризацию.

Пусть  $v^T$  — строка  $C$ , которую нужно удалить. Без потери общности можно считать, что  $v^T$  — последняя строка  $C$ . Пусть  $\tilde{C}$  — подматрица, образованная прочими строками  $C$ :

$$C = \begin{bmatrix} \tilde{C} \\ v^T \end{bmatrix}. \quad (27.34)$$

Заметим, что ранг  $\tilde{C}$  равен  $k$  либо  $k - 1$ . Практические приложения операции удаления данных чаще всего связаны только со случаем, когда

$\text{rank } C = \text{rank } \tilde{C} = n$ . Мы рассмотрим, однако, общий случай

$$\text{rank } C - 1 \leq \text{rank } \tilde{C} \leq \text{rank } C = k \leq n$$

в той мере, в какой он не потребует значительных дополнительных умозаключений.

Нужно найти множитель Холесского  $\tilde{R}$  для  $\tilde{C}$ , т.е. верхнюю треугольную  $k \times n$ -матрицу того же ранга, что и  $\tilde{C}$ , для которой

$$\tilde{Q}\tilde{C} = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix}, \quad (27.35)$$

где  $\tilde{Q}$  — некоторая ортогональная матрица порядка  $m - 1$ . Матрица  $\tilde{R}$  удовлетворяет еще соотношениям

$$\tilde{R}^T \tilde{R} = \tilde{C}^T \tilde{C} = C^T C - vv^T = R^T R - vv^T. \quad (27.36)$$

Первые два метода, которые мы опишем, были предложены в [68].

1. Удаление строки. В этом методе предполагается, что известна не только матрица  $R$ , но и матрица  $Q$  из (27.32). Представим  $Q$  в блочном виде и перепишем формулу (27.32):

$$\begin{matrix} k \\ 1 \\ m-k-1 \end{matrix} \left\{ \begin{matrix} Q_1 & p \\ u^T & \alpha \\ Q_2 & q \end{matrix} \right\} \cdot \begin{bmatrix} \tilde{C} \\ v^T \end{bmatrix} = \begin{bmatrix} R \\ 0 \\ 0 \end{bmatrix} \quad (27.37)$$

Умножим обе части (27.37) слева на ортогональную матрицу, которая преобразует вектор  $q$  в нулевой, изменяет  $\alpha$  и не меняет первые  $k$  строк  $Q$ . В качестве такой матрицы можно взять, например, одно преобразование Хаусхолдера или произведение  $m - k - 1$  преобразований Гивенса. Равенство (27.37) приобретает вид

$$\begin{bmatrix} Q_1 & p \\ \hat{u}^T & \hat{\alpha} \\ \hat{Q}_2 & 0 \end{bmatrix} \cdot \begin{bmatrix} \tilde{C} \\ v^T \end{bmatrix} = \begin{bmatrix} R \\ 0 \\ 0 \end{bmatrix} \quad (27.38)$$

Теперь умножим обе части (27.38) слева на последовательность  $k$  преобразований Гивенса, которые постепенно аннулируют элементы  $p$ , меняя при этом  $\hat{\alpha}$ . Пусть  $G_{ij}$  обозначает матрицу Гивенса, оперирующую со строками  $i$  и  $j$ . Первое из левых умножений будет на  $G_{k,k+1}$ , второе — на  $G_{k-1,k+1}$  и т.д. Последнее умножение будет на  $G_{1,k+1}$ . Эта последовательность операций обеспечивает, что преобразованная матрица  $R$ , которую мы обозначим  $\bar{R}$ , сохраняет верхнюю треугольную форму.

После всех преобразований равенство (27.38) переходит в

$$\begin{bmatrix} \bar{Q}_1 & 0 \\ \bar{u}^T & \bar{\alpha} \\ \hat{Q}_2 & 0 \end{bmatrix} \cdot \begin{bmatrix} \tilde{C} \\ v^T \end{bmatrix} = \begin{bmatrix} \bar{R} \\ \bar{w}^T \\ 0 \end{bmatrix}. \quad (27.39)$$

Так как  $\bar{\alpha}$  — единственный ненулевой элемент своего столбца, а столбец

имеет единичную евклидову длину, то  $\bar{\alpha} = \pm 1$ . Поскольку евклидовы длины строк также равны единице, то  $\bar{u}$  должен быть нулевым вектором. Это значит, что  $\bar{w}^T = \bar{\alpha} v^T = \pm v^T$ . Поэтому (27.39) можно переписать в виде

$$\begin{bmatrix} \bar{Q}_1 & 0 \\ 0 & \pm 1 \\ \bar{Q}_2 & 0 \end{bmatrix} \cdot \begin{bmatrix} \bar{C} \\ v^T \end{bmatrix} = \begin{bmatrix} \bar{R} \\ \pm v^T \\ 0 \end{bmatrix}. \quad (27.40)$$

Положим

$$\bar{Q} = \begin{bmatrix} \bar{Q}_1 \\ \bar{Q}_2 \end{bmatrix}. \quad (27.41)$$

Матрица  $\bar{Q}$  — ортогональная матрица порядка  $m-1$ ,  $\bar{R}$  — верхняя треугольная  $k \times n$ -матрица и

$$\bar{Q}\bar{C} = \begin{bmatrix} \bar{R} \\ 0 \end{bmatrix}. \quad (27.42)$$

Тем самым  $\bar{Q}$  и  $\bar{R}$  удовлетворяют требованиям, предъявляемым к матрицам  $\bar{Q}$  и  $\bar{R}$  (см (27.35) и (27.36)).

Каков ранг  $R$  ( $k$  или  $k-1$ ), зависит от того, равно или нет нулю число  $\hat{\alpha}$  в (27.38). По предположению все  $k$  диагональных элементов  $R$  ненулевые. Если  $\hat{\alpha} \neq 0$ , то, исследуя структуру и действие матриц Гивенса, участвующих в переходе от (27.38) к (27.40), легко убеждаемся, что все  $k$  диагональных элементов  $\bar{R}$  также будут ненулевыми.

Пусть, напротив,  $\hat{\alpha} = 0$ . Тогда  $\|p\| = 1$ , и, следовательно, некоторые компоненты вектора  $p$  отличны от нуля. Пусть  $p_l$  — последняя ненулевая компонента  $p$ , т.е.  $p_l \neq 0$ , и если  $l \neq k$ , то  $p_i = 0$  для  $i < l \leq k$ . Учитывая порядок, в котором применяются матрицы  $G_{ij}$  при переходе от (27.38) к (27.40), видим, что  $G_{l,k+1}$  будет первой матрицей, не совпадающей с единичной. Эта матрица  $G_{l,k+1}$  (с точностью до знака) является матрицей перестановки строк  $l$  и  $k+1$ , возможно, с изменением знака одной из этих строк. В частности, ее действие на  $R$  состоит в замене строки  $l$  строкой нулевых элементов. Последующие преобразования не меняют эту строку. Следовательно, в (27.40)  $l$ -я строка  $\bar{R}$  нулевая.

Итак, в этом случае  $\text{rank } \bar{R}$  будет меньше, чем  $k$ . С другой стороны,  $\text{rank } \bar{R}$  не может быть меньше  $k-1$ , так как  $\text{rank } \bar{R} = \text{rank } \bar{C} \geq \text{rank } C - 1 = k-1$ .

2. Удаление строки. В задаче, где  $m$  очень велико или используется последовательное накапливание, матрицу  $Q$  из (27.32) обычно не сохраняют. В этом случае метод 1 нельзя применить непосредственно. Однако все, что нужно от  $Q$  при определении матрицы  $\bar{R}$  в методе 1, — это вектор  $p$  из (27.37) и (27.38) и число  $\hat{\alpha}$  из (27.38). Мы увидим, что эти величины  $p$  и  $\hat{\alpha}$  можно вычислить по  $R$  и  $v$ . Перепишем (27.32) в виде

$$C = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix}. \quad (27.43)$$

Используя блочное представление (27.37), находим

$$v^T = p^T R \quad (27.44)$$

или, что то же самое,

$$R^T p = v. \quad (27.45)$$

Если  $R$  имеет ранг  $n$ , то (27.45) — это невырожденная треугольная система, которую можно разрешить относительно  $p$ .

Если  $k = \text{rank } R < n$ , то система (27.45) все же совместна и имеет единственное  $k$ -мерное решение  $p$ . По предположению первые  $k$  строк  $R^T$  составляют треугольную  $k \times k$ -подматрицу с ненулевыми диагональными элементами. Эта подматрица вместе с первыми  $k$  компонентами вектора  $v$  определяет систему уравнений, которую можно разрешить относительно  $p$ .

Определив  $p$ , можно вычислить число  $\hat{\alpha}$  по формуле

$$\hat{\alpha} = (1 - \|p\|^2)^{1/2}. \quad (27.46)$$

Эта формула является следствием того, что в (27.38) евклидова длина столбца  $(p^T, \hat{\alpha}, 0)^T$  равна единице. Величина  $\hat{\alpha}$  имеет произвольный знак и может быть взята неотрицательной, что и сделано в (27.46).

После того как  $p$  и  $\hat{\alpha}$  вычислены, матрицу  $\bar{R}$  можно построить с помощью  $k$  преобразований Гивенса, как это объяснено в изложении метода 1 вслед за формулой (27.38). Эту часть процесса можно описать равенством

$$G_{1,k+1} G_{2,k+1} \dots G_{k,k+1} \begin{bmatrix} p & R \\ \hat{\alpha} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \bar{R} \\ \pm 1 & \pm w^T \end{bmatrix}. \quad (27.47)$$

Следует ожидать, что метод 2 будет несколько менее точной численной процедурой, чем метод 1, из-за необходимости вычислять  $p$  и  $\hat{\alpha}$  по формулам (27.45) и (27.46). На точность, с которой определяется вектор  $p$ , влияет число обусловленности  $R$ ; оно, разумеется, равно числу обусловленности  $C$ .

Это ограничение на точность является, по-видимому, неизбежным в любом методе модификации множителя Холецкого  $R$  при удалении данных, который не использует дополнительных хранимых матриц типа  $Q$  или  $C$ . Можно ожидать, что при заданной точности машинной арифметики методы модификации, оперирующие непосредственно с матрицей  $(A^T A)^{-1}$ , будут значительно менее надежными, чем метод 2.

3. Удаление строки. Этот метод интересен тем, что требует меньше операций, чем метод 2, и алгоритмически очень схож с одним из методов, используемых при введении дополнительных данных. Это позволяет построить очень компактную программу, выполняющую и введение, и удаление информации. Сравнительная численная надежность методов 2 и 3 не исследовалась.

Метод 3 рассматривался в [37, 65, 66, 71]. Следуя примеру авторов этих работ, мы ограничимся обсуждением случая, когда  $\text{rank } \bar{C} = \text{rank } C = k = n$ . Свойства более общего случая выясняются в упражнениях.

Пусть  $i$  обозначает мнимую единицу, т.е.  $i^2 = -1$ . Равенство (27.36) можно переписать в виде

$$\tilde{R}^T \tilde{R} = \begin{bmatrix} R \\ i v^T \end{bmatrix}^T \cdot \begin{bmatrix} R \\ i v^T \end{bmatrix}. \quad (27.48)$$

Формально используя соотношения Гивенса (3.5), (3.8) и (3.9), можем построить матрицы  $F^{(l)}$ , которые приводят к треугольному виду матрицу  $\begin{bmatrix} R \\ iv^T \end{bmatrix}$ . Это достигается следующим процессом:

$$\begin{bmatrix} R^{(0)} \\ iv^{(0)T} \end{bmatrix} = \begin{bmatrix} R \\ iv^T \end{bmatrix}, \quad (27.49)$$

$$\begin{bmatrix} R^{(l)} \\ iv^{(l)T} \end{bmatrix} = F^{(l)} \begin{bmatrix} R^{(l-1)} \\ iv^{(l-1)T} \end{bmatrix}, \quad l = 1, \dots, k, \quad (27.50)$$

$$\begin{bmatrix} \bar{R} \\ 0 \end{bmatrix} = \begin{bmatrix} R^{(k)} \\ iv^{(k)T} \end{bmatrix}. \quad (27.51)$$

Здесь матрица  $F^{(l)}$  отличается от единичной только элементами, стоящими на пересечении строк  $l$  и  $k+1$  со столбцами  $l$  и  $k+1$ . В этих четырех позициях  $F^{(l)}$  находится подматрица

$$\begin{bmatrix} \sigma^{(l)} & i\tau^{(l)} \\ -i\tau^{(l)} & \sigma^{(l)} \end{bmatrix}, \quad (27.52)$$

где

$$\delta^{(l)} = [r_{ll}^{(l-1)}]^2 - [v_l^{(l-1)}]^2, \quad (27.53)$$

$$\rho^{(l)} = [\delta^{(l)}]^{1/2}, \quad (27.54)$$

$$\sigma^{(l)} = r_{ll}^{(l-1)} / \rho^{(l)}, \quad (27.55)$$

$$\tau^{(l)} = v_l^{(l-1)} / \rho^{(l)}. \quad (27.56)$$

Из нашего предположения о полноте ранга обеих матриц  $C$  и  $\tilde{C}$  можно вывести, что числа  $\delta^{(l)}$ , определяемые формулой (27.53), положительны.

Умножение на  $F^{(l)}$ , указанное в (27.50), можно выразить целиком в действительной арифметике:

$$r_{ij}^{(l)} = \sigma^{(l)} r_{ij}^{(l-1)} - \tau^{(l)} v_j^{(l-1)}, \quad (27.57)$$

$$v_j^{(l)} = -\tau^{(l)} r_{ij}^{(l-1)} + \sigma^{(l)} v_j^{(l-1)}, \quad j = l, \dots, n. \quad (27.58)$$

Способ построения  $F^{(l)}$  гарантирует выполнение условий  $r_{ll} = \rho^{(l)}$  и  $v_l^{(l)} = 0$ . Далее легко проверить, что  $v^{(k)} = 0$ , каждая матрица  $R^{(l)}$  верхняя треугольная,  $F^{(l)T} F^{(l)} = I$  и

$$\bar{R}^T \bar{R} = \begin{bmatrix} R \\ iv \end{bmatrix}^T \cdot \begin{bmatrix} R \\ iv \end{bmatrix}. \quad (27.59)$$

Тем самым  $\bar{R}$  удовлетворяет всем требованиям, предъявляемым к матрице  $\tilde{R}$ , т.е.  $\bar{R}$  есть верхний треугольный множитель Холесского для матрицы  $\tilde{C}$  из (27.34).

На практике потенциальная внутренняя неустойчивость задачи удаления данных может проявляться в потере значащих цифр при вычитании в формуле (27.53), а также в больших величинах множителей  $\sigma^{(l)}$  и  $\tau^{(l)}$  в формулах (27.57) и (27.58).

В работах Джентльмена [65, 66] отмечено, что модификации метода Гивенса, в которых выделен массив для хранения квадратов масштабирующих строчных множителей, особенно хорошо приспособиваются к методу 3 удаления строк. Простое допущение отрицательных значений для квадратов масштабирующих множителей позволяет представлять строки с чисто мнимым масштабированием, например  $iv^T$  в (27.48). Если для  $d_2$  в (10.31) и  $\tilde{d}_2$  в (10.33) допускаются не только положительные, но и отрицательные значения, то метод, описываемый формулами (10.31)–(10.44), сможет обрабатывать как приписывание, так и удаление данных. Заметим, что для отрицательных значений  $d_2$  величина  $l$  из (10.35) уже не обязана находиться в отрезке  $[0, 1]$ ; поэтому предотвращение машинных нулей и переполнений в  $\tilde{D}$  и  $\tilde{B}$  становится более сложной задачей.

У п р а ж н е н и я

27.60 [37]. а) Показать, что числа  $\sigma^{(l)}$  и  $\tau^{(l)}$  в формулах (27.55) и (27.56) можно интерпретировать соответственно как секанс и тангенс некоторого угла  $\theta^{(l)}$ .

б) Показать, что эти углы  $\theta^{(l)}$  совпадают с углами преобразований Гивенса  $G^{(l)}$  ( $c^{(l)} = \cos \theta^{(l)}$ ,  $s^{(l)} = \sin \theta^{(l)}$ ), которые участвуют в процессе приведения матрицы  $\bar{R}$ , окаймленной строкой  $v^T$ , к треугольному виду  $R$ ; этот процесс можно описать равенством

$$G^{(k)} \dots G^{(1)} \begin{bmatrix} \bar{R} \\ v^T \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

с) Показать, что при прежней трактовке чисел  $\theta^{(l)}$ ,  $c^{(l)}$  и  $s^{(l)}$  значение  $v_j^{(l)}$  из (27.58) можно вычислить по формуле

$$v_j^{(l)} = -s^{(l)} r_{ij}^{(l)} + c^{(l)} v_j^{(l-1)},$$

где  $c^{(l)} = 1/\sigma^{(l)}$  и  $s^{(l)} = \tau^{(l)}/\sigma^{(l)}$ .

27.61. Показать, что метод 3 удаления строк теоретически можно распространить на общий случай  $\text{rank } C - 1 < \text{rank } \tilde{C} < \text{rank } C \equiv k < n$ , опираясь на следующие замечания. (Мы продолжаем придерживаться заключенного в начале § 5 соглашения, что все  $k$  диагональных элементов  $R$  ненулевые.)

а) Если  $\delta^{(l)}$  неположительно для некоторых индексов  $l$  из последовательности  $1, \dots, k$ , то пусть  $h$  – первый такой индекс, т.е.  $\delta^{(h)} < 0$ , и если  $h \neq 1$ , то  $\delta^{(l)} > 0$  для  $1 \leq l < h$ . Показать, что  $\delta^{(h)} = 0$ .

б) Если индекс  $h$  определен, как в п. а), то  $v_h^{(h-1)} = \epsilon r_{hh}^{(h-1)}$ , где  $\epsilon = +1$  или  $\epsilon = -1$ . Показать, что в этом случае верны и соотношения  $v_j^{(h-1)} = \epsilon r_{hj}^{(h-1)}$ ,  $j = h+1, \dots, n$ .

с) При прежнем определении индекса  $h$  показать, что на шаге  $h$  алгоритм можно закончить, принимая в качестве конечной матрицы  $\bar{R}$  матрицу  $R^{(h-1)}$ , в которой строка  $h$  заменена нулевой строкой.

## ПРИЛОЖЕНИЕ А ОСНОВЫ ЛИНЕЙНОЙ АЛГЕБРЫ

В этом приложении мы перечислим основные используемые в книге факты линейной алгебры, не пытаясь представить логически завершенный набор понятий. Наша цель — кратко ввести лишь те понятия, которые прямо связаны с излагаемым в книге материалом.

Для действительного числа  $x$  положим

$$\operatorname{sgn} x = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$$

Под  $n$ -вектором  $x$  мы понимаем упорядоченный набор  $n$  (действительных) чисел  $x_1, \dots, x_n$ ;  $m \times n$ -матрица  $A$  — это прямоугольный массив (действительных) чисел, имеющий  $m$  строк и  $n$  столбцов. Элемент, стоящий на пересечении  $i$ -й строки и  $j$ -го столбца, обозначается через  $a_{ij}$ . Для  $m \times n$ -матрицы  $A$  мы часто используем символ  $A_{m \times n}$ .

Транспонированной для  $m \times n$ -матрицы  $A$  называется  $n \times m$ -матрица, обозначаемая через  $A^T$ , элемент которой, стоящий на пересечении  $i$ -й строки и  $j$ -го столбца, равен  $a_{ji}$ .

Произведение  $m \times n$ -матрицы  $A$  и  $l \times k$ -матрицы  $B$ , записываемое как  $AB$ , определено лишь при  $l = n$ . В этом случае  $C = AB$  есть  $m \times k$ -матрица с элементами

$$c_{ij} = a_{i1}b_{1j} + \dots + a_{in}b_{nj} = \sum_{p=1}^n a_{ip}b_{pj}.$$

Часто бывает удобно рассматривать  $n$ -вектор как  $n \times 1$ -матрицу. В частности, таким путем можно определить произведение матрицы на вектор.

Часто бывает удобно, напротив, считать  $m \times n$ -матрицу составленной из  $m$   $n$ -векторов — ее строк — или  $n$   $m$ -векторов — ее столбцов.

Скалярное произведение (называемое также внутренним произведением) двух  $n$ -мерных векторов  $u$  и  $v$  определяется формулой

$$s = u^T v = \sum_{i=1}^n u_i v_i.$$

Два вектора ортогональны друг другу, если их скалярное произведение равно нулю.

Евклидова длина, или евклидова норма, или  $l_2$ -норма вектора  $v$  обозначается через  $\|v\|$  и определяется как

$$\|v\| = (v^T v)^{1/2} = \left( \sum_{i=1}^n v_i^2 \right)^{1/2}.$$

Эта норма удовлетворяет неравенству треугольника

$$\|u + v\| \leq \|u\| + \|v\|$$

и обладает свойствами положительной однородности

$$\|\alpha u\| = |\alpha| \|u\|$$

(здесь  $\alpha$  — число, а  $u$  — вектор) и положительной определенности:

$$\|u\| > 0, \text{ если } u \neq 0, \quad \|u\| = 0, \text{ если } u = 0.$$

Эти три свойства характеризуют абстрактное понятие нормы.

Спектральная норма матрицы  $A$  обозначается через  $\|A\|$  и определяется как

$$\|A\| = \max \{ \|Av\| : \|v\| = 1 \}. \quad (\text{A.1})$$

Нам часто придется использовать и норму Фробениуса (называемую также нормой Шура или евклидовой матричной нормой) матрицы  $A$ , которая обозначается через  $\|A\|_F$  и определяется формулой

$$\|A\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 \right)^{1/2}. \quad (\text{A.2})$$

Спектральная норма и норма Фробениуса удовлетворяют соотношениям

$$\max_{i,j} |a_{ij}| \leq \|A\| \leq \|A\|_F \leq k^{1/2} \|A\|, \quad (\text{A.3})$$

где  $A$  —  $m \times n$ -матрица, а  $k = \min(m, n)$ .

Обе названные нормы удовлетворяют трем свойствам абстрактной нормы и, кроме того, мультипликативным неравенствам

$$\|AB\| \leq \|A\| \|B\|, \quad \|AB\|_F \leq \|A\|_F \|B\|_F.$$

Символ 0 будет обозначать нулевой вектор или нулевую матрицу; что именно и размерность величины определяются контекстом.

Система векторов  $v_1, \dots, v_k$  называется линейно зависимой, если существуют скаляры  $\alpha_1, \dots, \alpha_k$ , не все равные нулю и такие, что

$$\sum_{i=1}^k \alpha_i v_i = 0. \quad (\text{A.4})$$

Обратно, если условие (A.4) выполняется лишь при  $\alpha_1 = \dots = \alpha_k = 0$ , то векторы линейно независимы.

Множество всех  $n$ -мерных векторов образует  $n$ -мерное векторное пространство. Заметим, что если  $u$  и  $v$  — элементы этого векторного пространства, то это же верно для  $u + v$  и  $\alpha v$ , где  $\alpha$  — произвольный скаляр. Указанные два условия замкнутости относительно сложения векторов и умножения вектора на число характеризуют абстрактное определение векторного пространства. Однако нам придется иметь дело только с конкретным конечномерным векторным пространством, элементами которого являются наборы из  $n$  действительных чисел.

Если подмножество  $T$  векторного пространства  $S$  замкнуто относительно сложения векторов и умножения вектора на число, то  $T$  называется подпространством. Существует максимальное число линейно независимых векторов в подпространстве  $T$ . Это число  $m$  есть размерность подпространства  $T$ . Максимальная система линейно независимых векторов подпространства  $T$  называется базисом  $T$ . Каждое подпространство  $T$  размерности

$m \geq 1$  имеет базис. При этом для любой системы из  $k$ ,  $k < m$ , линейно независимых векторов  $m$ -мерного подпространства  $T$  найдутся в  $T$   $m - k$  добавочных вектора такие, что вместе эти  $m$  векторов составляют базис  $T$ . Если векторы  $u_1, \dots, u_m$  образуют базис  $T$  и  $v \in T$ , то существует единственный набор скаляров  $\alpha_i$ , для которого  $v = \sum_{i=1}^m \alpha_i u_i$ .

Оболочкой системы векторов  $v_1, \dots, v_k$  называется множество всех линейных комбинаций этих векторов, т.е. множество всех векторов вида  $u = \sum_{i=1}^k \alpha_i v_i$ , для произвольных скаляров  $\alpha_i$ . Оболочка системы из  $k$  векторов является подпространством размерности  $m$ ,  $m \leq k$ .

Некоторые подпространства возникают естественным образом в связи с матрицами. Так, с  $m \times n$ -матрицей  $A$  мы связываем образ, или пространство столбцов, т.е. оболочку ее столбцов; нулевое пространство, или ядро, т.е. множество  $\{x: Ax = 0\}$ , и пространство строк, т.е. оболочку строк. Заметим, что пространство строк  $A$  совпадает с образом  $A^T$ .

Часто бывает полезно следующее замечание: образ произведения матриц, скажем  $A = UVW$ , есть подпространство образа самого левого сомножителя произведения — в данном случае  $U$ . Точно так же пространство строк  $A$  есть подпространство пространства строк самого правого сомножителя — в данном случае  $W$ .

Пространства строк и столбцов матрицы  $A$  имеют одинаковую размерность. Это число называется рангом матрицы  $A$  и обозначается через  $\text{rank } A$ . Матрица  $A_{m \times n}$  имеет неполный ранг, если  $\text{rank } A < \min(m, n)$ , и полный ранг, если  $\text{rank } A = \min(m, n)$ . Квадратная матрица  $A_{n \times n}$  не вырождена, если  $\text{rank } A = n$ , и вырождена, если  $\text{rank } A < n$ .

Вектор  $v$  ортогонален к подпространству  $T$ , если  $v$  ортогонален к каждому вектору из  $T$ . Для ортогональности к  $T$  достаточно, чтобы  $v$  был ортогонален к каждому вектору некоторого базиса  $T$ . Подпространство  $T$  ортогонально к подпространству  $U$ , если любые векторы  $t \in T$  и  $u \in U$  ортогональны друг другу. Если подпространства  $T$  и  $U$  ортогональны, то прямой суммой  $T$  и  $U$  называется подпространство, обозначаемое через  $V = T \oplus U$  и определяемое как

$$V = \{v: v = t + u, t \in T, u \in U\}.$$

Размерность  $V$  равна сумме размерностей  $T$  и  $U$ .

Если  $T$  и  $U$  — взаимно ортогональные подпространства  $n$ -мерного векторного пространства  $S$  и  $S = T \oplus U$ , то  $T$  и  $U$  называются ортогональными дополнениями друг друга. Это записывается так:  $T = U^\perp$  и  $U = T^\perp$ . Для любого подпространства  $T$  существует ортогональное дополнение  $T^\perp$ . Если  $T$  — подпространство  $S$  и  $s \in S$ , то найдутся единственные векторы  $t \in T$  и  $u \in T^\perp$  такие, что  $s = t + u$ . Для этих векторов выполняется условие Пифагора:  $\|s\|^2 = \|t\|^2 + \|u\|^2$ .

Линейное многообразие — это сдвиг подпространства, т.е. если  $T$  — подпространство  $S$  и  $s \in S$ , то множество  $L = \{v: v = s + t, t \in T\}$  есть линейное многообразие. Размерность линейного многообразия определяется как размерность (единственного) ассоциированного с ним под-

пространства  $T$ . Гиперплоскость  $H$  в  $n$ -мерном векторном пространстве  $S$  — это  $n - 1$ -мерное линейное многообразие.

Если  $H$  — гиперплоскость и  $h_0 \in H$ , то  $T = \{t: t = h - h_0, h \in H\}$  —  $n - 1$ -мерное подпространство, а  $T^\perp$  одномерно. Если  $u \in T^\perp$ , то  $u^T h$  имеет одно и то же значение (скажем,  $d$ ) для всех  $h \in H$ . Таким образом, для заданных вектора  $u$  и скаляра  $d$  гиперплоскость  $H$  можно охарактеризовать как множество  $\{x: u^T x = d\}$ . В практических вычислениях гиперплоскости появляются именно через характеризацию этого типа.

Полупространство — это множество векторов, лежащих по одну сторону от гиперплоскости, т.е. множество вида  $\{x: u^T x \geq d\}$ . Многогранником называется пересечение конечного числа полупространств, т.е. множество вида  $\{x: u_i^T x \geq d_i, i = 1, \dots, m\}$  или, что эквивалентно, вида  $\{x: Ux \geq d\}$ , где  $U$  —  $m \times n$ -матрица со строками  $u_i^T$ ,  $d$  —  $m$ -вектор с компонентами  $d_i$  и неравенство интерпретируется поэлементно.

Элементы  $a_{ij}$  матрицы  $A$  называются ее диагональными элементами. Множество всех элементов  $a_{ii}$  называется главной диагональю  $A$ . Если все прочие элементы  $A$  нулевые, то  $A$  — диагональная матрица. Если  $a_{ij} = 0$  для  $|i - j| > 1$ , то  $A$  — трехдиагональная матрица.

Если  $a_{ij} = 0$  для  $j < i$  и  $j > i + 1$ , то матрица  $A$  верхняя двухдиагональная. Если  $a_{ij} = 0$  для  $j < i$ , то матрица  $A$  верхняя треугольная. Если  $a_{ij} = 0$  для  $j < i - 1$ , то матрица  $A$  верхняя хессенбергова. Если  $A^T$  — верхняя двухдиагональная, треугольная или хессенбергова матрица, то  $A$  — соответственно нижняя двухдиагональная, треугольная или хессенбергова матрица.

Квадратная диагональная  $n \times n$ -матрица, все диагональные элементы которой равны единице, называется единичной матрицей и обозначается  $I_n$  или просто  $I$ .

Если  $BA = I$ , то матрица  $B$  левая обратная для  $A$ . Матрица  $A_{m \times n}$  имеет левую обратную матрицу тогда и только тогда, когда  $\text{rank } A = n \leq m$ . Левая обратная матрица единственна в том и только том случае, если  $\text{rank } A = n = m$ . Операция транспонирования позволяет аналогичным образом определить правую обратную матрицу.

Если матрица  $A$  квадратная и невырожденная, то существует матрица, обозначаемая  $A^{-1}$ , которая одновременно является единственной левой обратной и единственной правой обратной для  $A$ . Она называется обратной матрицей для  $A$ .

Обобщением понятия обратной матрицы является псевдообратная матрица, которую обозначают через  $A^+$ . Псевдообратная матрица однозначно определена для любой  $m \times n$ -матрицы  $A$  и совпадает с обычной обратной матрицей, если матрица  $A$  квадратная и невырожденная. Пространства строк и столбцов у  $A^+$  те же, что и у  $A^T$ . Понятие псевдообратной матрицы, тесно связанное с задачей наименьших квадратов, определяется и обсуждается в гл. 7.

Квадратная матрица  $Q$  называется ортогональной, если  $Q^T Q = I$ . Из единственности обратной матрицы следует, что и  $Q Q^T = I$ .

Система векторов называется ортонормальной, если векторы этой системы попарно ортогональны и имеют единичную евклидову длину. Ясно, что система столбцов ортогональной матрицы ортонормальная, и такова же система ее строк.



Если  $Q_{m \times n}$ ,  $m \geq n$ , имеет ортонормальные столбцы, то  $\|Q\| = 1$ ,  $\|Q\|_F = n^{1/2}$ ,  $\|QA\| = \|A\|$  и  $\|QA\|_F = \|A\|_F$ . Если  $Q_{m \times n}$ ,  $m < n$ , имеет ортонормальные строки, то  $\|Q\| = 1$ ,  $\|Q\|_F = m^{1/2}$ ,  $\|AQ\| = \|A\|$  и  $\|AQ\|_F = \|A\|_F$ . Заметим, что ортогональная матрица  $Q_{n \times n}$  удовлетворяет обоим этим наборам условий.

Некоторые конкретные ортогональные матрицы, полезные в вычислениях, — это матрица вращения (Гивенса)

$$\begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix},$$

матрица отражения (Гивенса)

$$\begin{bmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{bmatrix}$$

и матрица отражения (Хаусхолдера)

$$H = I - 2 \frac{uu^T}{\|u\|^2},$$

где  $u$  — произвольный ненулевой вектор.

Матрица перестановки — это квадратная матрица, столбцы которой получаются перестановкой столбцов единичной матрицы. Матрица перестановки ортогональна.

Квадратная матрица  $A$  называется симметричной, если  $A^T = A$ . Симметричная матрица имеет спектральное разложение вида

$$A = QEQ^T,$$

где  $Q$  — ортогональная матрица, а  $E$  — (действительная) диагональная матрица. Диагональные элементы  $E$  суть собственные значения  $A$ , а столбцы  $Q$  — собственные векторы  $A$ . При этом  $j$ -й столбец  $q_j$  матрицы  $Q$  относится к  $j$ -му собственному значению  $e_{jj}$  и выполнено равенство  $Aq_j = e_{jj}q_j$ . Матрица  $A - e_{jj}I$  вырождена. Собственные значения симметричной  $n \times n$ -матрицы  $A$  определены однозначно. Если среди  $n$  собственных значений  $A$  число  $\lambda$  встречается  $m$  раз, то однозначно определено  $m$ -мерное подпространство, натянутое на  $m$  собственных векторов (столбцов  $Q$ ), относящихся к  $\lambda$ . Оно называется собственным подпространством  $A$ , относящимся к собственному значению  $\lambda$ .

Симметричная матрица положительно определена, если все ее собственные значения положительны. Положительно определенная матрица  $P$  характеризуется также тем свойством, что  $x^T Px > 0$  для всех  $x \neq 0$ .

Симметричная матрица  $S$  неотрицательно определена, если все ее собственные значения неотрицательны. Такая матрица обладает свойством  $x^T Sx \geq 0$  для всех  $x \neq 0$ .

Для произвольной  $m \times n$ -матрицы  $A$  матрица  $S = A^T A$  симметрична и неотрицательно определена. Она положительно определена, если  $\text{rang } A = n$ .

Инвариантное подпространство квадратной матрицы  $A$  — это такое подпространство  $T$ , что из  $x \in T$  следует  $Ax \in T$ . Если  $S$  — симметричная матрица, то каждое инвариантное подпространство  $S$  есть оболочка некоторой системы собственных векторов  $S$ , и, наоборот, оболочка любого множества собственных векторов  $S$  является инвариантным подпространством для  $S$ .

Симметричная матрица  $P$  называется проекционной матрицей, если все ее собственные значения равны единице или нулю. Матрица  $P$  называется идемпотентой, если  $P^2 = P$  (эквивалентно  $P(I - P) = 0$ ). Матрица  $P$  является проекционной матрицей в том и только том случае, если  $P$  — симметричная идемпотента.

Пусть  $P_{n \times n}$  — проекционная матрица,  $k$  собственных значений которой равны единице. Пусть  $T$  есть  $k$ -мерное собственное подпространство  $P$ , относящееся к собственному значению единица. Подпространство  $T$  — это (единственное) подпространство, ассоциированное с проекционной матрицей  $P$ , а  $P$  — (единственная) проекционная матрица, ассоциированная с подпространством  $T$ . Подпространство  $T$  является одновременно пространством строк и пространством столбцов для ассоциированной с ним проекционной матрицы  $P$  и характеризуется свойством

$$T = \{x: Px = x\}.$$

Если  $P$  — проекционная матрица с ассоциированным подпространством  $T$  и  $\|Px\| = \|x\|$ , то  $Px = x$  и  $x \in T$ . Кроме того,  $\|Px\| < \|x\|$  для  $x \notin T$ , и, следовательно,  $\|P\| = 1$ , если  $P \neq 0$ .

Для произвольной матрицы  $A_{m \times n}$  матрица  $A^+ A$  есть проекционная  $n \times n$ -матрица, ассоциированная с пространством строк  $A$ ;  $(I - A^+ A)$  — проекционная матрица, ассоциированная с ядром  $A$ ;  $AA^+$  — проекционная матрица, ассоциированная с образом (пространством столбцов)  $A$ .

Для любого  $n$ -вектора  $x$  существует единственное представление вида  $x = t + u$ , где  $t \in T$ , а  $u \in T^\perp$ . При этом  $t = Px$ ,  $u = (I - P)x$ . Вектор  $t$  — это ближайший к  $x$  вектор в  $T$  в том смысле, что  $\|x - t\| = \min \{\|x - v\|: v \in T\}$ . Вектор  $t$  называется проекцией  $x$  на  $T$ .

Обсуждение свойств проекционных операторов и доказательство большей части высказанных выше утверждений о них читатель найдет в книге [12].

Произвольная матрица  $A_{m \times n}$  допускает сингулярное разложение  $A = U_{m \times m} S_{m \times n} V_{n \times n}^T$ , где  $U$  и  $V$  — ортогональные матрицы, а  $S$  — диагональная матрица с неотрицательными диагональными элементами. Диагональные элементы  $S$  ( $s_{ii}$ ,  $i = 1, \dots, k$ , где  $k = \min(m, n)$ ) называются сингулярными числами  $A$ . Это множество чисел однозначно определяется матрицей  $A$ . Число ненулевых сингулярных чисел равно рангу  $A$ . Кроме того,

$$\|A\| = \max_i \{s_{ii}\}, \quad \|A\|_F = \left( \sum_{i=1}^k s_{ii}^2 \right)^{1/2},$$

что дает простое доказательство двух правых неравенств (A.3).

Пусть  $A = USV^T$  — сингулярное разложение матрицы  $A_{m \times n}$ . Тогда спектральными разложениями матриц  $A^T A$  и  $AA^T$  будут соответственно  $A^T A = V(S^T S) V^T$  и  $AA^T = U(SS^T) U^T$ .

Сингулярное разложение дает информацию, полезную при практическом анализе задач линейной алгебры. Его свойства и приложения обсуждаются в гл. 4, 5, 18, 25, 26.

# ПРИЛОЖЕНИЕ В

## ДОКАЗАТЕЛЬСТВО ГЛОБАЛЬНОЙ КВАДРАТИЧНОЙ СХОДИМОСТИ QR-АЛГОРИТМА

Назначение этого приложения — дать доказательство теоремы 18.5. Приводимое нами доказательство является более подробным вариантом того, что содержится в [194, 195]. Утверждение а) теоремы 18.5 вынесено в упражнение 18.46.

Лемму В.1 вместе с доказательством можно найти в книге [7].

**Лемма В.1.** Пусть  $A$  — симметричная трехдиагональная  $n \times n$ -матрица с диагональными элементами  $a_1, \dots, a_n$  и наддиагональными (поддиагональными) элементами  $b_2, \dots, b_n$ , причем все  $b_i$  ненулевые. В таком случае  $A$  имеет  $n$  различных собственных значений.

Нам понадобится также следующая простая лемма, доказательство которой вынесено в упр. В.52. Мы используем обозначения, введенные в тексте и уравнениях гл. 18 (от начала и до (18.4)).

**Лемма В.2.** Для всех  $k = 1, 2, \dots$  элементы матриц  $A_k$  и сдвиги  $\sigma_k$  ограничены по модулю величиной  $\|A\|$ , а элементы матриц  $A_k - \sigma_k I_n$  и  $R_k$  ограничены по модулю величиной  $2\|A\|$ .

Мы должны исследовать основные операции QR-алгоритма со сдвигами, чтобы установить свойства некоторых промежуточных величин и в конечном счете оценить величину некоторых внедиагональных элементов матриц  $A_k$ .

Обозначим диагональные элементы сдвинутой матрицы  $A_k - \sigma_k I_n$  через

$$\bar{a}_i^{(k)} = a_i^{(k)} - \sigma_k,$$

$$i = 1, \dots, n.$$

Согласно правилу выбора  $\sigma_k$  (см. (18.4)), то собственное значение нижней угловой  $2 \times 2$ -подматрицы в матрице  $(A_k - \sigma_k I_n)$ , которое ближе к  $\bar{a}_n^{(k)}$ , равно нулю. Поэтому

$$\bar{a}_n^{(k)} \bar{a}_{n-1}^{(k)} = (b_n^{(k)})^2, \quad (B.3)$$

$$|\bar{a}_n^{(k)}| \leq |b_n^{(k)}| \leq |\bar{a}_{n-1}^{(k)}|. \quad (B.4)$$

Ортогональная матрица  $Q_k$  представляется произведением

$$Q_k = J_{n-1,n}^{(k)} \dots J_{2,3}^{(k)} J_{1,2}^{(k)}, \quad (B.5)$$

где каждая матрица  $J_{i-1,i}^{(k)}$  есть вращение в плоскости, определяемой  $i-1$ -й

и  $i$ -й координатными осями. Она имеет вид

$$J_{i-1,i}^{(k)} = \begin{bmatrix} I_{i-2} & 0 & 0 \\ 0 & P_i^{(k)} & 0 \\ 0 & 0 & I_{n-i} \end{bmatrix}, \quad i = 2, \dots, n, \quad (B.6)$$

где  $P_i^{(k)}$  — матрица двумерного вращения, определяемая, как и в (3.5), соответствующими числами  $c_i^{(k)}$  и  $s_i^{(k)}$ .

После того как матрица  $A_k - \sigma_k I_n$  умножена слева на первые  $i-2$  матриц вращений, имеем

$$J_{i-2,i-1}^{(k)} \dots J_{1,2}^{(k)} (A_k - \sigma_k I_n) =$$

$$= \begin{bmatrix} p_1^{(k)} & q_1^{(k)} & r_1^{(k)} & & & \\ 0 & p_2^{(k)} & q_2^{(k)} & r_2^{(k)} & & \\ & & & & & \\ & & & p_{i-2}^{(k)} & q_{i-2}^{(k)} & r_{i-2}^{(k)} \\ & & & 0 & x_{i-1}^{(k)} & y_{i-1}^{(k)} \\ & & & & b_i^{(k)} & \bar{a}_i^{(k)} & b_{i+1}^{(k)} \\ & & & & & & b_{n-1}^{(k)} & \bar{a}_{n-1}^{(k)} & b_n^{(k)} \\ 0 & & & & & & 0 & b_n^{(k)} & \bar{a}_n^{(k)} \end{bmatrix} \quad (B.7)$$

Умножая обе части (B.7) слева на  $J_{i-1,i}^{(k)}$ , выводим, используя формулы (3.7) — (3.9), следующие рекуррентные соотношения:

$$p_{i-1}^{(k)} = [(x_{i-1}^{(k)})^2 + (b_i^{(k)})^2]^{1/2}, \quad i = 2, \dots, n, \quad (B.8)$$

$$p_n^{(k)} = x_n^{(k)}, \quad (B.9)$$

$$c_i^{(k)} = \frac{x_{i-1}^{(k)}}{p_{i-1}^{(k)}}, \quad i = 2, \dots, n, \quad (B.10)$$

$$s_i^{(k)} = \frac{b_i^{(k)}}{p_{i-1}^{(k)}}, \quad i = 2, \dots, n, \quad (B.11)$$

$$0 = -s_i^{(k)} x_{i-1}^{(k)} + c_i^{(k)} b_i^{(k)}, \quad i = 2, \dots, n, \quad (B.12)$$

$$x_i^{(k)} = -s_i^{(k)} y_{i-1}^{(k)} + c_i^{(k)} \bar{a}_i^{(k)}, \quad i = 2, \dots, n, \quad (B.13)$$

$$y_i^{(k)} = c_i^{(k)} b_{i+1}^{(k)}, \quad i = 2, \dots, n-1. \quad (B.14)$$

Подобное преобразование завершается построением матрицы

$$A_{k+1} = R_k Q_k^T + \sigma_k I_n = R_k (J_{1,2}^{(k)})^T \dots (J_{n-1,n}^{(k)})^T + \sigma_k I_n. \quad (B.15)$$

В этом процессе новые внедиагональные элементы вычисляются по формулам

$$b_i^{(k+1)} = s_i^{(k)} p_i^{(k)}, \quad i = 2, \dots, n. \quad (B.16)$$

Мы хотим проанализировать поведение последовательности  $\{b_n^{(k+1)}\}$  при  $k \rightarrow \infty$ .

В формулах (B.17)–(B.19) мы опустим верхний индекс  $k$ , но сохраним индекс  $k+1$ . Начиная с (B.16) при  $i = n$ , имеем

$$\begin{aligned} b_n^{(k+1)} &= s_n p_n = s_n x_n = s_n [-s_n y_{n-1} + c_n \bar{a}_n] = \\ &= s_n \left[ -s_n c_{n-1} b_n + s_n \left( \frac{\bar{a}_n}{b_n} \right) x_{n-1} \right] = \\ &= s_n \left[ -s_n c_{n-1} b_n + s_n \left( \frac{b_n}{\bar{a}_{n-1}} \right) (-s_{n-1} c_{n-2} b_{n-1} + c_{n-1} \bar{a}_{n-1}) \right] = \\ &= -\frac{s_n^2 s_{n-1} c_{n-2} b_{n-1} b_n}{\bar{a}_{n-1}}. \end{aligned} \quad (B.17)$$

Во втором из этих преобразований использована формула (B.9), затем последовательно формулы (B.13), (B.14) и (B.12) и, наконец, в предпоследнем переходе – формулы (B.3), (B.13) и (B.14).

Исключая  $p_{n-1}$  из (B.11) и (B.16), находим

$$b_{n-1}^{(k+1)} = \frac{s_{n-1} b_n}{s_n}. \quad (B.18)$$

Теперь из (B.17) и (B.18) получаем

$$b_n^{(k+1)} b_{n-1}^{(k+1)} = -s_n s_{n-1}^2 c_{n-2} b_n b_{n-1} \left( \frac{b_n}{\bar{a}_{n-1}} \right). \quad (B.19)$$

Из формулы (B.4), записанной в виде  $|b_n^{(k)} / \bar{a}_{n-1}^{(k)}| \leq 1$ , и формулы (B.17) выводим

$$|b_n^{(k+1)}| \leq |b_n^{(k)}|, \quad k = 1, 2, \dots \quad (B.20)$$

Из (B.19) следует, что

$$|b_n^{(k+1)} b_{n-1}^{(k+1)}| \leq |b_n^{(k)} b_{n-1}^{(k)}|, \quad k = 1, 2, \dots, \quad (B.21)$$

поэтому последовательность  $|b_n^{(k)} b_{n-1}^{(k)}|$  сходится при  $k \rightarrow \infty$  к пределу  $L \geq 0$ .

**Лемма B.22.** Предел  $L$  равен нулю.

**Доказательство.** Предположим, что  $L > 0$ . Переходя к пределу при  $k \rightarrow \infty$  в обеих частях равенства (B.19), получаем

$$|b_n^{(k)} / \bar{a}_{n-1}^{(k)}| \rightarrow 1, \quad (B.23)$$

$$|s_n^{(k)}| \rightarrow 1, \quad (B.24)$$

$$|c_{n-2}^{(k)}| \rightarrow 1, \quad (B.25)$$

$$(s_{n-1}^{(k)})^2 \rightarrow 1. \quad (B.26)$$

Из (B.10), (B.24) и ограниченности последовательности  $\{p_{n-1}^{(k)}\}$  (см. лемму B.2) выводим, что

$$x_{n-1}^{(k)} \rightarrow 0. \quad (B.27)$$

Поэтому, согласно (B.13) и (B.14),

$$\bar{a}_{n-1}^{(k)} c_{n-1}^{(k)} - b_{n-1}^{(k)} c_{n-2}^{(k)} s_{n-1}^{(k)} \rightarrow 0. \quad (B.28)$$

Так как (B.26) подразумевает, что  $c_{n-1}^{(k)} \rightarrow 0$ , то из (B.28) следует

$$b_{n-1}^{(k)} c_{n-2}^{(k)} s_{n-1}^{(k)} \rightarrow 0. \quad (B.29)$$

Из (B.25), (B.26) и (B.29) вытекает, что  $b_{n-1}^{(k)} \rightarrow 0$ . Поскольку последовательность  $\{b_n^{(k)}\}$  ограничена, отсюда следует  $|b_{n-1}^{(k)} b_n^{(k)}| \rightarrow 0$ , что противоречит предположению  $L > 0$ . Это доказывает лемму B.22.

**Лемма B.30.** Последовательность  $\{|b_n^{(k)}|\}$ ,  $k = 1, 2, \dots$ , содержит произвольно малые члены.

**Доказательство.** Фиксируем число  $\tau > 0$ . По лемме B.2 последовательности  $\{|b_n^{(k)}|\}$  и  $\{|b_{n-1}^{(k)}|\}$  ограничены. Поэтому, согласно лемме B.22, найдется целое число  $\bar{k}$ , для которого либо

$$|b_{n-1}^{(\bar{k})}| < \tau, \quad (B.31)$$

либо

$$|b_n^{(\bar{k})}| < \tau. \quad (B.32)$$

Если выполняется (B.31), то из (B.20) выводим  $|b_n^{(\bar{k}+1)}| < \tau$ . Итак, для любого  $\tau > 0$  существует номер  $\hat{k}$ , зависящий от  $\tau$ , для которого  $|b_n^{(\hat{k})}| < \tau$ . Лемма (B.30) доказана.

**Лемма B.33.** Если  $\tau > 0$  достаточно мало, а номер  $\hat{k}$  таков, что  $|b_n^{(\hat{k})}| < \tau$ , то  $|b_n^{(k)}| < \tau$  для всех  $k > \hat{k}$ .

**Доказательство.** Пусть  $\delta = \min \{|\lambda_i - \lambda_j| : i \neq j\}$ . Заметим, что  $\delta > 0$ , так как собственные значения  $\lambda_i$  матрицы  $A$  различны. Положим

$$f(t) = \frac{t [1 + t/(\delta - 3t)]}{\delta - 3t}. \quad (B.34)$$

Пусть  $\tau_0 > 0$  настолько мало, чтобы выполнялись условия:

$$\tau_0 < \delta/3, \quad (B.35)$$

$$f(\tau_0) < 1, \quad (B.36)$$

$$\frac{df(t)}{dt} > 0, \quad 0 \leq t \leq \tau_0. \quad (B.37)$$

Выберем число  $\tau$  в интервале  $(0, \tau_0)$ , и пусть  $k$  — целое число, для которого

$$|b_n^{(k)}| < \tau. \quad (B.38)$$

Удобно будет ввести обозначение

$$\epsilon = b_n^{(k)}. \quad (B.39)$$

Если  $b_n^{(k)} = 0$ , то алгоритм сошелся. В противном случае без ограничения общности можем считать, что  $\epsilon > 0$ .

Так как  $\bar{a}_n^{(k)} \bar{a}_{n-1}^{(k)} = \epsilon^2$ , можно написать

$$A_k - \sigma_k I_n = \left[ \begin{array}{ccc|c} & & & 0 \\ & B & & \dots \\ & & & 0 \\ \hline & & & \epsilon \\ \hline 0 & \dots & 0 & \epsilon \end{array} \right] \quad (B.40)$$

Пусть  $\mu_1, \dots, \mu_{n-1}$  — собственные значения симметричной матрицы  $B$  порядка  $n-1$ , а  $\lambda'_1, \dots, \lambda'_n$  — собственные значения матрицы  $A_k - \sigma_k I_n$ . По теореме 5.1 упорядоченные собственные значения матрицы  $A_k - \sigma_k I_n$  и собственные значения матрицы

$$\left[ \begin{array}{cc} B & 0 \\ 0 & \bar{a}_{n-1}^{(k)} \end{array} \right]$$

различаются не более, чем на  $\epsilon$ . Поэтому выполняются неравенства

$$|\lambda'_i - \mu_i| \leq \epsilon, \quad i = 1, \dots, n-1, \quad (B.41)$$

$$\left| \lambda'_n - \frac{\epsilon^2}{\bar{a}_{n-1}^{(k)}} \right| \leq \epsilon,$$

где числа  $\lambda'_i$ , возможно, подверглись переиндексации. Из тождества

$$\mu_i = \mu_i - \lambda'_i + \lambda'_i - \lambda'_n + \lambda'_n - \frac{\epsilon^2}{\bar{a}_{n-1}^{(k)}} + \frac{\epsilon^2}{\bar{a}_{n-1}^{(k)}}$$

следует

$$|\mu_i| \geq |\lambda'_i - \lambda'_n| - |\mu_i - \lambda'_i| - \left| \lambda'_n - \frac{\epsilon^2}{\bar{a}_{n-1}^{(k)}} \right| - \frac{\epsilon^2}{|\bar{a}_{n-1}^{(k)}|}, \quad i = 1, \dots, n-1. \quad (B.42)$$

Применяя теперь (B.41) для второго и третьего членов в правой части неравенства (B.42), замечая, что  $|\lambda'_i - \lambda'_n| = |\lambda_i - \lambda_n| \geq \delta$ , и, наконец, используя для четвертого члена формулу (B.4) вместе с  $b_n^{(k)} = \epsilon$ , имеем

$$|\mu_i| \geq \delta - 3\epsilon, \quad i = 1, \dots, n-1. \quad (B.43)$$

Перед последним шагом приведения матрицы  $A_k - \sigma_k I_n$  к верхнему треугольному виду нижняя угловая  $2 \times 2$ -подматрица устроена так:

$$\begin{bmatrix} x_{n-1}^{(k)} & \epsilon c_{n-1}^{(k)} \\ \epsilon & \epsilon^2 / \bar{a}_{n-1}^{(k)} \end{bmatrix}. \quad (B.44)$$

Это следует из (B.3), (B.7), (B.14) и (B.39).

Заметим, что в формуле (B.44)  $x_{n-1}^{(k)}$  — диагональный элемент верхней треугольной матрицы порядка  $n-1$ , полученной в результате левого умножения матрицы  $B$  из (B.40) на последовательность из  $n-2$  матриц вращения. Поэтому, согласно (6.3) и (B.43),

$$|x_{n-1}^{(k)}| \geq \min |\mu_i| \geq \delta - 3\epsilon. \quad (B.45)$$

По завершении подобного преобразования из (B.3), (B.9), (B.13), (B.17) и равенств  $y_{n-1}^{(k)} = c_{n-1}^{(k)} b_n^{(k)} = c_{n-1}^{(k)} \epsilon$  выводим

$$\begin{aligned} b_n^{(k+1)} &= s_n^{(k)} p_n^{(k)} = (c_n^{(k)} \bar{a}_n^{(k)} - s_n^{(k)} y_{n-1}^{(k)}) s_n^{(k)} = \\ &= \left( \frac{\epsilon^2 c_n^{(k)}}{\bar{a}_{n-1}^{(k)}} - \epsilon c_{n-1}^{(k)} s_n^{(k)} \right) s_n^{(k)}. \end{aligned} \quad (B.46)$$

Теперь из (B.11) и неравенства (B.45) следует

$$|s_n^{(k)}| = \frac{\epsilon}{[(x_{n-1}^{(k)})^2 + \epsilon^2]^{1/2}} \leq \frac{\epsilon}{\delta - 3\epsilon}. \quad (B.47)$$

Наконец, согласно (B.46) и (B.47),

$$|b_n^{(k+1)}| \leq \frac{\epsilon^3}{|\bar{a}_{n-1}^{(k)}| (\delta - 3\epsilon)} + \frac{\epsilon^3}{(\delta - 3\epsilon)^2}. \quad (B.48)$$

Неравенство (B.48) показывает, что если последовательность  $\{|\bar{a}_{n-1}^{(k)}|\}$  ограничена снизу положительным числом, то сходимость асимптотически будет кубической. Однако в общем случае из (B.4) следует лишь, что

11

(B. 49)

$$|b_n^{(k+1)}| \leq [b_n^{(k)}]^2 \frac{1 + b_n^{(k)} / (\delta - 3b_n^{(k)})}{\delta - 3b_n^{(k)}} = b_n^{(k)} f(b_n^{(k)}). \quad (\text{B. 50})$$
$$|b_n^{(k+1)}| < |b_n^{(k)}|. \quad (\text{B.51})$$

3. Доказательство, которое мы привели, относится к случаю  $n \geq 3$ .  
Если  $n = 2$  и

$$A = A_1 = \begin{bmatrix} a_1^{(1)} & b_2^{(1)} \\ b_2^{(1)} & a_2^{(1)} \end{bmatrix},$$

$$A_2 = \begin{bmatrix} a_1^{(2)} & b_2^{(2)} \\ b_2^{(2)} & a_3^{(2)} \end{bmatrix},$$

4. Более общо, если вместо  $\sigma_k$  из (18.4) в качестве сдвига используется точное собственное значение  $\lambda$  матрицы  $A$ , то матрица  $A_{k+1}$  распадается

**У п р а ж н е н и е**  
В. 52. Доказать лемму В. 2.

[illegible]

SECRET

U.S. GOVERNMENT PRINTING OFFICE: 1967 O - 380-000

FOR SALE BY THE NATIONAL ARCHIVES AT COLLEGE PARK, MARYLAND

U.S. GOVERNMENT PRINTING OFFICE: 1967 O - 380-000

FOR SALE BY THE NATIONAL ARCHIVES AT COLLEGE PARK, MARYLAND

*[Faint handwritten notes at the bottom of the page]*

[illegible]

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED  
DATE 08-11-2010 BY 60322 UCBAW/SJS/STP

## ПОСЛЕСЛОВИЕ

Х.Д. Икрамов

Ниже приведен комментарий к ряду глав прочитанной вами книги. В нем кратко прореферированы важнейшие работы, относящиеся к сквозной теме книги — прямым методам для линейных задач наименьших квадратов. В основном это публикации 1974 г. и более поздних лет, и их отсутствие в книге понятно. В то же время в нескольких случаях работы, включенные в список литературы книги, в основном тексте не упоминаются. Так произошло, например, с циклом статей Бьорка по итерационному уточнению и работой Пейджа [50\*]. Комментарий заканчивается собственным списком литературы, и ссылки здесь даются к этому дополнительному списку.

Глава 4. Понятия сингулярного числа и сингулярного разложения перенесены в [64\*] на случай пары матриц с одинаковым числом столбцов; при этом строчные размеры могут не совпадать. Пусть  $A$  —  $m_1 \times n$ -матрица, а  $B$  —  $m_2 \times n$ -матрица. Если матричный пучок  $A^T A - \lambda B^T B$  несингулярен, т.е. характеристический многочлен  $\det(A^T A - \lambda B^T B)$  не равен нулю тождественно по  $\lambda$ , то конечные собственные значения пучка неотрицательны. Арифметические корни из них называются (обобщенными) сингулярными числами пары матриц  $(A, B)$ . Если  $B = I_n$  — единичная  $n \times n$ -матрица, то приходим к стандартному определению сингулярных чисел.

Предположим (имея в виду приложения к анализу задач с ограничениями), что  $m_1 \geq n$ . Тогда [64\*] существуют ортогональные  $m_1 \times m_1$ -матрица  $U$  и  $m_2 \times m_2$ -матрица  $V$ , а также невырожденная  $n \times n$ -матрица  $X$  такие, что \*

$$U^T A X = D_A = \text{diag}(\alpha_1, \dots, \alpha_n), \alpha_i \geq 0, \quad (1)$$

$$V^T B X = D_B = \text{diag}(\beta_1, \dots, \beta_s), \beta_i \geq 0, s = \min(m_2, n), \quad (2)$$

$$\beta_1 \geq \dots \geq \beta_r > 0, \beta_{r+1} = \dots = \beta_s = 0, r = \text{rang } B. \quad (3)$$

Соотношения (1), (2) вместе с условиями (3) называются (обобщенным) сингулярным разложением пары  $(A, B)$ . При этом отношения  $\alpha_i/\beta_i$ ,  $i = 1, \dots, r$ , суть сингулярные числа этой пары.

Остроумное доказательство этого факта, упрощающее доказательство Ван Лоана и опирающееся на (неявно) введенное еще в [25] понятие

\* Запись  $\text{diag}(\gamma_1, \dots, \gamma_t)$ ,  $t = \min(m, n)$ , обозначает  $m \times n$ -матрицу, у которой элементы главной диагонали равны числам  $\gamma_i$ , а все прочие элементы равны нулю.

CS-разложения ортонормальной матрицы, дано в [55]. Предположим для простоты, что ранг  $(m_1 + m_2) \times n$ -матрицы

$$F = \begin{bmatrix} A \\ B \end{bmatrix}$$

равен  $n$ . Пусть

$$F = QR = \begin{bmatrix} Q_A \\ Q_B \end{bmatrix} R \quad (4)$$

есть ортогонально-треугольное разложение  $F$ , где  $R$  — невырожденная верхняя треугольная матрица, а блочное разбиение ортонормальной  $(m_1 + m_2) \times n$ -матрицы  $Q$  индуцировано аналогичным разбиением  $F$ . Если  $Q_A = UCW^T$  — сингулярное разложение  $Q_A$ , то в матрице

$$G = \begin{bmatrix} U^T & 0 \\ 0 & I_{m_2} \end{bmatrix} QW = \begin{bmatrix} C \\ Y \end{bmatrix} \quad (5)$$

верхний блок диагональный. Сингулярное разложение неоднозначно, и мы будем считать, что  $C$  выбрана так, что сингулярные числа, равные единице (если таковые имеются), занимают последние диагональные позиции  $C$ :

$$C = \begin{bmatrix} \tilde{C} & 0 \\ 0 & I_t \\ 0 & 0 \end{bmatrix} \}_{m_1 - n}$$

Таким образом,  $0 \leq \tilde{c}_{ii} < 1$ ,  $i = 1, 2, \dots, n - t$ .

Поскольку столбцы  $G$  имеют единичную длину, последние  $t$  столбцов  $Y$  должны быть нулевыми:

$$Y = \begin{bmatrix} \tilde{Y} \\ 0 \end{bmatrix} \}_{n-t \atop t}$$

Так как  $G^T G = I_n$ , то, в частности,  $\tilde{C}^T \tilde{C} + \tilde{Y}^T \tilde{Y} = I_{n-t}$ , т.е. матрица  $\tilde{Y}^T \tilde{Y}$  диагональная, и, следовательно, столбцы  $\tilde{Y}$  попарно ортогональны. Поэтому  $\tilde{Y}$  можно представить в виде  $\tilde{Y} = \tilde{V} \tilde{S}$ , где  $\tilde{V}$  —  $m_2 \times (n - t)$ -матрица с ортонормированными столбцами,  $\tilde{S}$  — диагональная матрица порядка  $n - t$ , диагональными элементами которой служат длины столбцов  $\tilde{Y}$ . Из сказанного заодно вытекает, что  $n - t = r = \text{rang } B$ .

Дополняя матрицу  $\tilde{V}$  до ортогональной матрицы  $V$  порядка  $m_2$ , а также окаймляя  $\tilde{S}$  нулевыми строками и столбцами до  $m_2 \times n$ -матрицы  $S$ , получим  $Y = VS$ . Это позволяет переписать (5) в виде

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix} QW = \begin{bmatrix} C \\ S \end{bmatrix} \quad (6)$$

$$\begin{bmatrix} U^T & 0 \\ 0 & V^T \end{bmatrix} \cdot \begin{bmatrix} A \\ B \end{bmatrix} X = \begin{bmatrix} C \\ S \end{bmatrix},$$

где  $X = R^{-1}W$ . Это и есть обобщенное сингулярное разложение пары  $(A, B)$ , причем  $C = D_A$ ,  $S = D_B$ .

Способ устойчивого вычисления CS-разложения (6), использующий модификацию метода вращений Якоби, описан в [61\*].

Сингулярное разложение пары матриц оказалось полезным средством исследования задач с ограничениями. Некоторые другие его приложения, а также другой способ обобщения понятия сингулярного числа указаны в [64\*].

Анализ чувствительности обобщенных сингулярных чисел и сомножителей обобщенного сингулярного разложения к возмущениям элементов данной матричной пары проведен в [39\*]. В частности, получены "обобщенные" варианты классических теорем Вейля—Лидского и Хоффмана—Виланда.

Г л а в а 5. Более точный по сравнению с теоремой 5.7 анализ возмущений сингулярных чисел выполнен в [60\*]. В обозначениях гл. 5 справедливо равенство

$$\beta_i^2 = (\alpha_i + \xi_i)^2 + \eta_i^2, \quad i = 1, \dots, k, \quad (7)$$

причем

$$|\xi_i| \leq \|PE\|, \quad \inf_{\|x\|=1} \|(I-P)Ex\| \leq \eta_i \leq \|(I-P)E\|.$$

Здесь  $P$  — ортогональный проектор (проекционная матрица) на образ  $A$ .

Если  $\alpha_i$  велико относительно  $\|E\|$ , то вторым квадратом в (7) можно пренебречь, т.е.  $\beta_i \approx \alpha_i + \xi_i$ . Поскольку

$$|\xi_i| \leq \|PE\| \leq \|E\|, \quad (8)$$

то теорема 5.7 содержится в (7). Однако (8) сильнее, поскольку  $\|PE\|$  может оказаться существенно меньше, чем  $\|E\|$ . В особенности это вероятно при  $k \ll m$ , т.е. когда размерность образа  $A$  много меньше размерности  $m$  пространства.

Для малых  $\alpha_i$ , когда оба квадрата в (7) сравнимы по величине, второе слагаемое  $\eta_i^2$  имеет тенденцию увеличивать значение сингулярного числа. Поэтому для плохо обусловленных матриц малые возмущения чаще приводят к улучшению обусловленности, а не к ее ухудшению.

В [62\*] получены точные выражения для членов первого и второго порядка в асимптотическом разложении младшего сингулярного числа возмущенной матрицы.

Г л а в а 6. Пусть вычислено ортогонально-треугольное разложение  $m \times n$ -матрицы  $A$  из задачи НК, причем у треугольной  $n \times n$ -матрицы  $R$  все диагональные элементы не равны нулю. Следует ли отсюда, что в пределах той точности, с какой ведутся вычисления,  $R$  не вырождена? Дадим эквивалентные постановки этого вопроса. Насколько мало младшее сингулярное число  $R$ ? Насколько велика норма матрицы  $R^{-1}$ , т.е. если считать  $R$  нормированной, насколько велико ее число обусловленности?

Теорема 6.13 показывает, что даже для треугольной матрицы того типа, что строится в алгоритме HFTI (т.е., в частности, имеющей монотонно убывающие диагональные элементы), младший диагональный элемент может не давать правильного представления о величине наименьшего сингулярного числа. В [41\*] предложен метод вычисления верхней оценки для  $\|R^{-1}\|_F$ , требующий  $O(n)$  операций. Он основан на том, что для треугольной матрицы  $\rho$  с теми же диагональными элементами, что и у  $R$  (которые в алгоритме HFTI всегда положительны), и отрицательными наддиагональными элементами, определяемыми формулой

$$\rho_{ij} = -a_i = -\max_{i < j \leq n} |r_{ij}|, \quad i = 1, \dots, n-1,$$

справедливо неравенство

$$\|R^{-1}\|_F \leq \|\rho^{-1}\|_F.$$

При этом

$$\|\rho^{-1}\|_F^2 = \sum_{i=1}^n \mu_i / r_{ii}^2,$$

а коэффициенты  $\mu_i$  вычисляются рекурсивно:

$$\mu_1 = 1,$$

$$\mu_i = (1 + c_{i-1})^2 \mu_{i-1} - 2c_{i-1}, \quad i = 2, 3, \dots, n;$$

$$c_i = a_i / r_{ii}, \quad i = 1, 2, \dots, n-1.$$

Многие современные программы решения линейных систем содержат блок оценки числа обусловленности матрицы. В случае треугольной матрицы  $R$  такой блок реализует один шаг обратной итерации для  $R^T R$  при специально выбираемом начальном приближении [21\*, 22\*].

Г л а в а 7. Существует необозримая литература о разных типах обобщенных обратных матриц. Так, в посвященной этому предмету книге [49\*] библиография содержит 1775 названий.

Пусть  $I$  — произвольное подмножество множества  $\{1, 2, 3, 4\}$ . Матрица  $X$  называется обобщенной обратной для  $A$  типа  $I$  (и обозначается  $A^I$ ), если для  $X$  выполнены все условия Пенроуза с номерами из  $I$ . В частности,  $A^{\{1, 2, 3, 4\}} = A^+$ . Для нетривиальных же подмножеств  $I$  множества обобщенных обратных матриц типа  $I$  в общем случае бесконечны.

Запись многих фактов, связанных с задачами НК, не требует именно псевдообратной матрицы: ее могут заменять более слабые обобщенные обратные. Так, формулу из упражнения 7.22, описывающую все псевдо-решения задачи  $Ax \cong b$ , можно заменить на

$$x = A^{\{1, 3\}} b + (I - A^{\{1\}} A) y.$$

Псевдообратная матрица  $A^+$  введена в теоремах 7.1, 7.3 как матрица, дающая при фиксированной  $m \times n$ -матрице  $A$  и произвольном  $m$ -векторе  $b$  решение минимальной длины (= нормальное псевдорешение) задачи  $Ax \cong b$ . Можно поставить более общую задачу: найти

$$\min_{x \in F} \|x\|_L, \quad F = \{x \mid \|Ax - b\|_M = \min\}, \quad (9)$$

где  $\|\cdot\|_L, \|\cdot\|_M$  — эллипсоидальные полунормы

$$\|x\|_L = \|Lx\|, \quad \|y\|_M = \|My\|.$$

Эта задача рассматривалась рядом авторов, в частности в [28\*, 47\*].

Положим  $P = I - (MA)^+ MA$ , т.е.  $P$  — проектор на ядро  $MA$ . Общее решение (9) записывается формулой

$$x = (I - (LP)^+ L) (MA)^+ Mb + P(I - (LP)^+ LP)z, \quad (10)$$

где  $z$  — произвольный вектор из  $R^n$ . Решение задачи (9) будет единственным в том и только том случае, когда

$$\ker(MA) \cap \ker L = \{0\}. \quad (11)$$

Слагаемые формулы (10) ортогональны, поэтому, даже если (11) не выполнено, вектор

$$x = (I - (LP)^+ L) (MA)^+ Mb \quad (12)$$

является (уже однозначно определенным) решением (9), имеющим минимальную евклидову длину. Это мотивирует следующее определение.

Матрица  $A_{M,L}^+ = (I - (LP)^+ L) (MA)^+ M$  называется  $ML$ -взвешенной псевдообратной для  $A$  [28\*]. Формула (12) принимает вид

$$x = A_{M,L}^+ b.$$

Матрица  $A_{M,L}^+$  принадлежит к числу обобщенных обратных для  $A$  типа {2}. Ряд ее свойств обобщает соответствующие свойства псевдообратных матриц Мура—Пенроуза. Известно, например, что

$$A^+ = \lim_{\delta \rightarrow 0} (A^T A + \delta I)^{-1} A^T.$$

Этому отвечает такое свойство взвешенной псевдообратной матрицы:

$$\lim_{\lambda \rightarrow 0} B(\lambda) = \lim_{\lambda \rightarrow 0} [(MA)^T MA + \lambda^2 L^T L]^+ (MA)^T M = A_{M,L}^+. \quad (13)$$

Отметим еще формулу

$$\lim_{\lambda \rightarrow +\infty} B(\lambda) = (MAP_L)^+ M, \quad (14)$$

где  $P_L = I - L^+ L$ .

Подобно тому как сингулярное разложение матрицы  $A$  упрощает вычисление псевдообратной матрицы (см. формулу (7.11)), обобщенное сингулярное разложение пары  $(A, B)$  позволяет определить взвешенные псевдообратные  $A_{I,B}^+, B_{I,A}^+$ . Именно в обозначениях формул (1), (2) справедливо:

$$A_{I,B}^+ = XD_A^+ U^T, \quad B_{I,A}^+ = XD_B^+ V^T.$$

**Глава 9.** Вопрос о том, насколько реалистичны оценки возмущений нормального псевдорешения из теоремы 9.7, разбирается в [59\*]. При этом наряду с возмущениями исходных данных такого типа, как в названной теореме, т.е. характеризуемых наличием малого положительного  $\epsilon$ , для которого

$$\|E\| < \epsilon \|A\|, \quad \|db\| < \epsilon \|b\|,$$

рассматриваются и возмущения, вносящие малую относительную погреш-

ность в каждый столбец  $A$  и в правую часть задачи. Если  $E_j(A_j)$  обозначает  $j$ -й столбец  $E(A)$ , то возмущения этого типа описываются неравенствами

$$\|E_j\| < \epsilon \|A_j\|, \quad j = 1, \dots, n, \quad \|db\| < \epsilon \|b\|.$$

Будем говорить соответственно о возмущениях класса I и класса II.

Приведем несколько важных результатов этой работы. Пусть  $s_1 \geq s_2 \geq \dots \geq s_n > 0$  — сингулярные числа  $m \times n$ -матрицы  $A$ , и пусть  $\mu = \epsilon \kappa = \epsilon \|A\| \|A^+\| < 1$ . Найдутся возмущения  $E_1, (db)_1$  класса I такие, что

$$\|dx\| \geq \frac{\epsilon}{s_n} \left[ \frac{s_1 \|r\|}{s_n(1-\mu^2)} + \frac{\|b\|}{1-\mu^2} \right],$$

и возмущения  $E_2, (db)_2$  класса I, для которых

$$\|dx\| \geq \frac{\epsilon}{s_n} [s_1 \|x\| + \|b\|].$$

Учитывая, что в рассматриваемом случае — матрица  $A$  имеет полный столбцовый ранг — четвертый член в правой части (9.9) (см. теорему 9.7) отсутствует, видим, что остальные три члена действительно должны входить в любую верхнюю оценку возмущений псевдорешения для возмущений исходных данных класса I.

В случае возмущений класса II ситуация более сложная. Положим  $\sigma = \|A\|_F$ ,  $\mu = \epsilon \sigma / s_n$ , и пусть  $\mu < 1$ . Для произвольной невырожденной диагональной  $n \times n$ -матрицы  $D$  положим  $\tilde{A} = AD$ ; пусть  $\tilde{\sigma}, \tilde{s}_n, \tilde{\mu}$  определены по отношению к  $\tilde{A}$  так же, как  $\sigma, s_n, \mu$  по отношению к  $A$ . Рассматриваем только такие матрицы  $D$ , для которых  $\tilde{\mu} < 1$ . Оказывается, что для любого возмущения класса II

$$\|dx\| \leq \frac{\epsilon}{s_n} \left[ \frac{\tilde{\sigma} \|r\|}{\tilde{s}_n(1-\tilde{\mu}^2)} + \frac{\sigma \|x\|}{1-\mu} + \frac{\|b\|}{1-\mu} \right]. \quad (15)$$

Главное отличие (15) от оценки (9.9) теоремы 9.7 заключается в том, что коэффициент  $\tilde{\sigma}/\tilde{s}_n$  при  $r$  заменяет прежний коэффициент  $s_1/s_n$ . Это значит, что если масштабированием столбцов можно существенно уменьшить число обусловленности  $A$  (выполнять это масштабирование на самом деле необязательно), то верхняя оценка для  $\|dx\|$  в действительности зависит лишь от первой степени  $\|A^+\|$ . Так будет, если столбцы  $A$  имеют сильно различающиеся длины, но хорошо разделенные направления.

Представляет интерес и следующее замечание из [59\*]. Будем рассматривать возмущения класса I. Введем угол  $\varphi$  между правой частью  $b$  и образом матрицы  $A$ . Существуют такие возмущения  $E, db$ , что для относительной погрешности решения выполняется нижняя оценка

$$\frac{\|dx\|}{\|x\|} \geq \epsilon \kappa^2 \operatorname{tg} \varphi \frac{1}{1-\mu^2}. \quad (16)$$

Предположим, что  $\mu = \epsilon \kappa \ll 1$ , но правая часть (16) больше 1. В этом случае вектор  $\tilde{x} = x + dx$  не имеет какой-либо цены как приближение к  $x$ . Тем



не менее он содержит в себе некоторую информацию об исходной задаче. Так, для не слишком малых  $\varphi$  можно показать, что  $r(\tilde{x}) \leq 1,005r$ , т.е. вектор  $\tilde{x}$  почти минимизирует длину невязки. Для углов  $\varphi$ , не слишком близких к  $\pi/2$ , вектор  $\tilde{x}$  позволяет определить проекцию  $b$  на образ  $A$  с малой относительной погрешностью. Это следует из приближенного равенства  $\|Adx\|/\|Ax\| \approx \mu \operatorname{tg} \varphi$ .

**Глава 13.** Относительная погрешность приближенного решения минимальной длины (= нормального решения), вычисленного алгоритмом НВТ — HS2, будет (в наихудшем случае) зависеть от первой степени числа обусловленности, а не от второй, как было бы для несовместной системы. Этот вывод можно сделать, проводя для алгоритма обратный анализ погрешностей округлений в духе гл. 15–17 и применяя теорему 9.18. К тому же заключению приводит и прямой анализ погрешностей [38\*].

Вместо алгоритма НВТ — HS2 нормальное решение можно искать так. Первым шагом, как и прежде, является ортогонально-треугольное разложение матрицы  $A$ :

$$AQ = [L : 0],$$

где  $L$  — нижняя треугольная матрица. Будем искать нормальное решение в виде

$$x = A^T w. \quad (17)$$

Подставляя (17) в систему  $Ax = b$ , получим

$$AA^T w = LL^T w = b. \quad (18)$$

Таким образом, вычисление  $x$  сводится к решению треугольных систем с матрицами  $L$  и  $L^T$ , после чего применяется (17). Этот подход позволяет не хранить матрицу  $Q$  ценой хранения  $A$ ; в больших разреженных задачах, где  $Q$  — значительно более плотная матрица, он дает существенную экономию памяти.

Можно думать, что, поскольку  $w$  определяется из системы (18), погрешность вычисленного вектора  $x$  может быть пропорциональна квадрату числа обусловленности к матрицы  $A$ . Как показано в [50\*], в действительности ситуация более благоприятна. В оценке погрешности, в самом деле, присутствует квадратичный по  $k$  член; он умножается, однако, на квадрат машинной точности  $\eta$ . Поэтому при  $k\eta \ll 1$  погрешность пропорциональна  $k\eta$ , а не  $k^2\eta$ .

В [23\*, 40\*] даны обзоры методов для недоопределенных систем. Во второй из этих статей типовая задача линейной теории упругости — расчет вектора сил при заданной конечно-элементной модели структуры и заданных нагрузках — интерпретируется как задача вычисления нормального решения недоопределенной системы полного строчного ранга. При этом используемые на практике методики расчета вектора сил — методы перемещений, сил, естественной факторизации — оказываются различными способами вычисления нормального решения.

**Глава 14.** Алгоритм НФТИ был независимо предложен Голубом [32\*] и — под названием нормализованного процесса — советским коллективом: Д.К. Фаддеев, В.Н. Кублановская, В.Н. Фаддеева [11\*]. Значение  $k$  псевдоранга в нормализованном процессе определяется, исходя

из величины диагональных элементов вычисляемой треугольной матрицы. Как отмечено в [26\*], обычный способ выбора  $k$  можно интерпретировать следующим образом: норма активной подматрицы становится сравнимой с нормой матрицы эквивалентного возмущения; при этом нормой  $m \times n$ -матрицы  $A$  служит величина

$$\|A\| = \max_{1 \leq k \leq n} \left( \sum_{i=1}^m a_{ik}^2 \right)^{1/2}.$$

Многочисленные приложения нормализованного процесса к различным задачам линейной алгебры указаны в работах В.Н. Кублановской (см., например, [5\*–8\*]).

**Главы 15–17.** Пусть  $A$  —  $m \times n$ -матрица псевдоранга  $n$ . Как следует из теорем 16.1, 17.11, приближенное решение  $\tilde{x}$  задачи  $Ax \cong b$ , вычисленное алгоритмами ортогонально-треугольного разложения, будет точным для задачи со слабо возмущенными входными данными  $A + E$ ,  $b + f$ . Однако в силу теоремы 9.7 это не означает, что разность  $dx = \tilde{x} - x$ , где  $x$  — точное решение задачи  $Ax \cong b$ , обязательно будет мала. Два фактора влияют на увеличение нормы  $dx$ : плохая обусловленность матрицы  $A$  и сильная несовместность системы (т.е. большая величина оптимальной невязки  $r$ ). Что же делать, если качество вычисленного приближения  $\tilde{x}$  неприемлемо?

При решении линейных систем с квадратными невырожденными матрицами в такой ситуации широко используется итерационное уточнение — метод, впервые предложенный и подробно исследованный в книге [66\*]. Техника итерационного уточнения была перенесена на случай решения задачи наименьших квадратов в работах [13\*, 32\*]. При этом использовался наиболее очевидный подход: если  $\delta$  — решение задачи НК  $A\delta \cong b - A\tilde{x}$ , то  $\tilde{x} + \delta$  есть решение исходной задачи  $Ax \cong b$ , т.е. совпадает с  $x$ . Таким образом, нужно хранить матрицу  $A$ , чтобы вычислить (обычно это делают с удвоенной точностью или посредством операции накопления) невязку  $\tilde{r} = b - A\tilde{x}$ , после чего решается задача НК с той же матрицей  $A$  и правой частью  $\tilde{r}$ . Если сохранено ортогонально-треугольное разложение  $A$ , решение новой задачи НК потребует лишь  $O(mn)$  операций вместо  $O(mn^2)$ . Если  $\tilde{\delta}$  — вычисленное решение этой задачи и точность приближения  $\tilde{\tilde{x}} = \tilde{x} + \tilde{\delta}$  все еще недостаточна, то процесс повторяется и т.д. Метод реализован алгольной программой [18\*].

Внутреннее ограничение, присущее этому подходу, вскрыто в [35\*]. Пусть на вход процедуры итерационного уточнения "подано" точное решение  $x$ , и пусть точно вычислена невязка  $r = b - Ax$ . Мы считаем, что  $r \neq 0$ . Точным решением задачи  $A\delta \cong r$  был бы нулевой вектор. Однако о вычисленном решении  $\tilde{\delta}$  мы можем сказать лишь, что оно является точным для задачи  $(A + E)\delta \cong r + f$ . Оценки теорем 16.1, 17.11, 9.7 показывают, что мы можем гарантировать для нормы  $\tilde{\delta}$  оценку типа

$$\|\tilde{\delta}\| \leq f(m, n) k^2 \|r\| / \|A\|. \quad (19)$$

Здесь  $f(m, n)$  — некоторое выражение от  $m$  и  $n$ , зависящее от деталей машинной арифметики и метода, в частности от того, используется или нет вычисление скалярных произведений с повышенной точностью. Чтобы сумма  $x + \tilde{\delta}$  не искажала точного решения, нужно, как минимум, потребо-

таким образом выполняется неравенство

$$\kappa^2 \|r\| / (\|A\| \|x\|) < 1. \quad (20)$$

Для задач с большими невязками и не очень хорошо обусловленной  $A$  это условие будет нарушено.

Пусть для решения задачи  $Ax \cong b$  строилось ортогонально-треугольное разложение  $A$ , и пусть  $\tilde{R}$  — вычисленная треугольная матрица этого разложения, а  $R$  — отвечающая ей точная матрица. Поскольку задача  $Ax \cong b$  теоретически эквивалентна системе нормальных уравнений  $A^T Ax = A^T b$  и  $A^T A = R^T R$ , то Кахан предложил следующую процедуру итерационного уточнения:

$$x^{(0)} = 0, \quad r^{(s)} = b - Ax^{(s)},$$

$$\tilde{R}^T \tilde{R} \delta^{(s)} = A^T r^{(s)}, \quad x^{(s+1)} = x^{(s)} + \delta^{(s)}, \quad s = 0, 1, 2, \dots$$

У этой процедуры, казалось бы, нет отмеченного выше недостатка процесса Голуба: если невязка  $r$  точно ортогональна образу  $A$ , то  $A^T r = 0$  и  $\delta = 0$ . Однако  $r^{(s)}$  и произведение  $A^T r^{(s)}$  не будут, вообще говоря, вычислены точно. Пусть  $x^{(s)}$  — правильно округленное точное решение  $x$ . Вместо точной невязки  $r$  в лучшем случае будет вычислена  $\tilde{r} = r + h$ , где  $\|h\| \leq \eta \|r\|$ . Так как  $A^T r = 0$ , то для  $\tilde{\delta}$  получаем систему

$$\tilde{R}^T \tilde{R} \tilde{\delta} = A^T h,$$

откуда

$$\|\tilde{\delta}\| \leq \kappa^2(A) \|r\| \eta / \|A\|,$$

т.е. практически то же самое, что и в (19).

Итак, для сильно несовместных систем описанные выше приемы итерационного уточнения работают неудовлетворительно. В работах Бюрка [14\*, 17\*] развит другой подход, использующий эквивалентность задачи  $Ax \cong b$  и системы линейных уравнений

$$B \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \cdot \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}, \quad (21)$$

где  $r = b - Ax$  — оптимальная невязка. Матрица  $B$  системы (21) квадратная, а если  $\text{rank } A = n$ , что в дальнейшем предполагается, то и невырожденная. Поэтому к (21) применим обычный процесс итерационного уточнения. Приводим его описание, следуя [14\*].

Полагаем  $r^{(0)} = 0, x^{(0)} = 0$ . Итерация с номером  $s$  состоит из трех шагов:

1. Вычислить невязки

$$\begin{bmatrix} f^{(s)} \\ g^{(s)} \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \cdot \begin{bmatrix} r^{(s)} \\ x^{(s)} \end{bmatrix}. \quad (22)$$

На этом шаге операции проводятся с удвоенной точностью.

2. Определить поправки  $\delta r^{(s)}, \delta x^{(s)}$ , решая систему

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \cdot \begin{bmatrix} \delta r^{(s)} \\ \delta x^{(s)} \end{bmatrix} = \begin{bmatrix} f^{(s)} \\ g^{(s)} \end{bmatrix}. \quad (23)$$

3. Найти новые приближения  $r^{(s+1)} = r^{(s)} + \delta r^{(s)}, x^{(s+1)} = x^{(s)} + \delta x^{(s)}$ .

Процедура Голуба получается, если в этом описании положить  $r^{(s)} = 0$  для всех  $s$ . Это означает, что в случае совместной системы асимптотическое поведение методов должно быть одинаковым, и также отчасти объясняет, почему при больших невязках процесс Голуба дает неудовлетворительные результаты.

Как решать систему (23)? Заметим, что при  $s = 0$  из формулы (22) следует, что  $f^{(0)} = b, g^{(0)} = 0$ , а тогда (23) превращается в (21). Таким образом, первая итерация есть попросту первоначальное решение исходной задачи  $Ax \cong b$ . Предположим, что для этого (методом НФТ или модифицированным методом Грама — Шмидта (MGS) из гл. 19) вычислено ортогонально-треугольное разложение  $A$  (мы рассматриваем пока случай точных вычислений):

$$QA = \begin{bmatrix} R \\ 0 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}. \quad (24)$$

Опуская в (23) верхний индекс, подставим в оба соотношения системы формулу (24), а затем умножим первое из них слева на  $Q$ . Тогда получим

$$Q\delta r + \begin{bmatrix} R \\ 0 \end{bmatrix} \delta x = Qf, \quad (25)$$

$$[R^T : 0] Q\delta r = g. \quad (26)$$

Представим вектор  $d = Qf$  в виде

$$d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} \begin{matrix} n \\ m-n \end{matrix}.$$

Из (26) следует, что верхние  $n$  компонент  $m$ -вектора  $Q\delta r$  образуют вектор  $h = (R^T)^{-1} g$ , а из (25) видно, что нижние  $m - n$  компонент  $Q\delta r$  составляют вектор  $d_2$ . Значит,

$$\delta r = Q^T \begin{bmatrix} h \\ d_2 \end{bmatrix}.$$

Вектор  $\delta x$  определяем, приравняв верхние компоненты в соотношении (25):

$$\delta x = R^{-1}(d_1 - h).$$

Итак, решение системы (23) требует: 1) вычисления вектора  $d^{(s)} = Qf^{(s)}$ ; 2) решения треугольной системы  $R^T h^{(s)} = g^{(s)}$ ; 3) вычисления

вектора

$$\delta r^{(s)} = Q^T \begin{bmatrix} h^{(s)} \\ d_2^{(s)} \end{bmatrix};$$

4) решения треугольной системы  $R\delta x^{(s)} = d_1^{(s)} - h^{(s)}$ . Для каждого из этих действий достаточно  $O(mn)$  или  $O(n^2)$  операций, что значительно меньше стоимости ортогонально-треугольного разложения  $A$ . Это же можно сказать об итерации процесса (шаги 1) – 3)) в целом.

В первой части работы [14\*] проведен очень тщательный анализ описанной процедуры итерационного уточнения для двух способов ортогонально-треугольного разложения: метода отражений и метода MGS. Главный вывод этого анализа: для не слишком плохо обусловленных задач процесс сходится в обоих случаях; сходимость линейная, и ее замедление при ухудшении обусловленности  $A$  пропорционально первой степени числа обусловленности. Это утверждение имеет место независимо от того, мала или нет оптимальная невязка. Для величины  $\lim_{s \rightarrow \infty} (\|x - x^{(s)}\| / \|x\|)$

можно дать верхнюю оценку, в которой главным членом является произведение

$$f(m, n) \kappa^2 \|r\| \eta^2 / (\|A\| \|x\|).$$

По сравнению с (19), (20) здесь прибавился множитель  $\eta$ . Поэтому, если выполнено (20), то процесс Бюрка приводит к результату, мало отличающемуся от округленного вектора  $x$ . Для случая, когда ортогонально-треугольное разложение вычисляется методом отражений, алгоритм программы процесса дана в [17\*].

При решении больших разреженных задач стараются избежать хранения ортогонального сомножителя в разложении матрицы  $A$ . Две модификации процесса Бюрка для таких задач описаны в [33\*, 29\*]. В первой из этих работ выводится следствие из (23), позволяющее вычислять  $\delta x^{(s)}$ , не используя  $Q$ . Именно

$$A\delta x^{(s)} = f^{(s)} - \delta r^{(s)}, \quad (27)$$

откуда, умножая обе части на  $A^T$  и используя равенство  $A^T \delta r^{(s)} = g^{(s)}$ , получаем

$$A^T A \delta x^{(s)} = A^T f^{(s)} - g^{(s)}.$$

Но  $A^T A = R^T R$ , поэтому окончательно находим

$$R^T R \delta x^{(s)} = A^T f^{(s)} - g^{(s)}. \quad (28)$$

Вычислив  $\delta x^{(s)}$  (т.е., согласно (28), построив произведение  $A^T f^{(s)}$ ), решив две треугольные системы), найдем  $\delta r^{(s)}$  из (27):

$$\delta r^{(s)} = f^{(s)} - A\delta x^{(s)}.$$

Шаги 1, 3 те же, что в методе Бюрка.

В качестве начального приближения этого процесса можно брать пару  $(x^{(0)}, r^{(0)})$ , где  $x^{(0)}$  – приближенное решение, вычисленное методом отражений;  $r^{(0)}$  – отвечающая ему невязка. Но, как отмечено в [15\*],

можно взять и  $x^{(0)} = r^{(0)} = 0$ . В этом случае  $Q$  не используется даже для формирования начального приближения, что является достоинством при необходимости решать серию задач с одной и той же матрицей  $A$  и разными правыми частями.

В [15\*] показано, что скорость сходимости этой модификации по-прежнему (как и в методе Бюрка) зависит от первой степени числа обусловленности  $A$ . Иначе обстоит дело во второй модификации [29\*], где в качестве матрицы  $R$  в (28) предлагается брать вычисленный методом Холесского треугольный множитель матрицы  $A^T A$ . Здесь (см. [15\*]) с ростом числа обусловленности  $A$  скорость сходимости падает пропорционально  $\kappa^2$ .

Итерационное уточнение для задачи неполного ранга анализируется в [2\*]. Для этого же случая во второй части работы [14\*] дана алгоритмная процедура.

Глава 18. Алгоритм Голуба–Бизингера–Райнша пришел на смену более раннему алгоритму сингулярного разложения, так называемому методу односторонней ортогонализации. Идея этого метода независимо высказана рядом авторов (см., например, [4\*]). Наиболее ранняя публикация принадлежит, по-видимому, Хестенсу [37\*], почему и говорят обычно об алгоритме Хестенса. В последние годы интерес к этому методу, одно время совершенно вытесненному из практики алгоритмом SVD, снова оживился в связи с удобством его реализации как на мини-ЭВМ, так и на мощных компьютерах четвертого поколения [46\*, 48\*].

Нужно отметить, что почти одновременно с Бизингером и Голубом и независимо от них алгоритм сингулярного разложения, очень схожий с алгоритмом SVD, построил В.В. Воеводина [1\*]. Основное различие обоих методов состоит в том, что шаг итерационной фазы в алгоритме Воеводина эквивалентен шагу QR-алгоритма без сдвигов для соответствующей трехдиагональной матрицы.

Высказанная в начале § 5 идея о том, что при больших размерах задачи и  $m \gg n$  целесообразно вначале привести матрицу к верхней треугольной форме последовательностью левых отражений, подробно развита в [19\*]. Этот вариант алгоритма сингулярного разложения становится эффективней стандартной процедуры уже при  $p = m/n > 2$ . Для  $p \approx 10$  экономия машинного времени может достигать 50%. Интересно и следующее наблюдение. Если для  $m = n$  любой вариант сингулярного разложения требует много большей работы в сравнении с нормализованным процессом, то уже при  $p \approx 3$  соотношение числа операций для модифицированного SVD и нормализованного процесса равно приблизительно 3. Это соотношение может оказаться вполне приемлемым, если учесть, что SVD дает значительно больше информации о задаче.

Там же, в [19\*], указана другая возможность экономии вычислений, правда, ценой введения дополнительного  $n \times n$ -массива. Если пользователю нужна матрица  $U$  левых сингулярных векторов, то обычно она вычисляется следующим образом. После завершения двухдиагонализации строится в явном виде произведение  $Q$  отражений  $Q_n \dots Q_1$  (точнее, первые  $n$  столбцов  $Q$ ); до этого каждая из матриц  $Q_i$  хранилась в компактной форме (т.е. фактически хранился соответствующий вектор  $u_i$ ) в нижней части массива  $A$ . Итак,  $Q$  –  $m \times n$ -матрица. В последующем каждое левое вращение  $T_j$  (по стандартной оценке в QR-процессе их  $\approx 2n^2$ ) применяется

к столбцам  $Q$ , требуя  $4m$  операций умножения (при обычной, "не быстрой" форме вращений). Более экономный способ состоит в накапливании произведения вращений  $T_i$  в виде  $n \times n$ -матрицы  $T$  в выделенном для этой цели дополнительном массиве. Лишь по окончании итерационной фазы алгоритма производится перемножение  $Q$  и  $T$ .

В [24\*] предлагается вообще отказаться от формирования в явном виде ортогональных сомножителей разложения; хранятся лишь порождающие векторы отражений и информация о вращениях. Если опорная пара индексов вращения известна, то для устойчивого восстановления  $s$  и  $z$  достаточно хранить одно число (см., например, [9\*, с. 110]). В  $QR$ -процессе вращения следуют одно за другим циклами  $(1, 2), (2, 3), \dots, (n-1, n)$ , пока матрица не будет расщеплена вследствие появления малого внедиагонального элемента. Показано, что информация о расщеплениях и, следовательно, полная информация об опорных индексах вращений может быть упакована в целом массиве длины  $n$ . Общие требования к памяти при таком методе хранения возрастают приблизительно в полтора раза, но и время вычисления сингулярного разложения сокращается по меньшей мере вдвое.

Отметим, что алгоритм программы алгоритма сингулярного разложения имеется в справочнике [10\*], а фортранная, как уже говорилось в предисловии, — в книге [12\*].

Глава 19. Метод Питерса—Уилкинсона, описанный в упражнении 19.38, привлек к себе внимание многих алгебраистов. Первые отклики [20\*, 56\*] представляли собой попытки усовершенствовать метод (в сторону уменьшения числа операций) в случае слабо переопределенных систем. Пусть  $L$  — нижняя треугольная  $m \times n$ -матрица, полученная на первом этапе метода Питерса—Уилкинсона. Для искомого решения  $x$  справедлива формула

$$x = R^{-1} (L^T L)^{-1} L^T P^T b. \quad (29)$$

Применим к  $L$  последовательность левых отражений  $Q_n, \dots, Q_1$ , определяемых следующим образом. Отражение  $Q_n$  действует на строки  $L$  с номерами  $n, n+1, \dots, m$  и аннулирует элементы  $(n+1, n), \dots, (m, n)$ ;  $Q_{n-1}$  действует на строки  $Q_n L$  с номерами  $n-1, n+1, \dots, m$  и аннулирует элементы  $(n+1, n-1), \dots, (m, n-1)$ ; наконец,  $Q_1$  действует на строки  $Q_2 \dots Q_n L$  с номерами  $1, n+1, \dots, m$  и аннулирует элементы  $(n+1, 1), \dots, (m, 1)$ . Хранить порождающие векторы отражений можно в освобожденных позициях  $m \times n$ -массива, первоначально содержащего матрицу  $L$ . Результатом описанного процесса будет нижняя треугольная  $n \times n$ -матрица  $\tilde{L}$ . Полагая  $Q = Q_1 \dots Q_n$ , имеем

$$QL = \begin{bmatrix} I_n \\ 0 \end{bmatrix} \tilde{L}. \quad (30)$$

Подставляя (30) в (29), получим

$$x = R^{-1} \tilde{L}^{-1} (I_n; 0) Q^T P^T b.$$

Это значит, что для вычисления  $x$  нужно найти первые  $n$  компонент вектора  $Q^T P^T b = Q_1 \dots Q_n P^T b$ , а затем решить две треугольные системы с матрицами  $\tilde{L}$  и  $R$ . Главный член числа операций умножения в изложенном методе равен  $3mn^2/2 - 7n^3/6$ . Это меньше, чем в исходном методе Питерса—Уил-

кинсона или в методе отражений, если  $m < 5n/3$ , и даже меньше, чем в обычном методе нормальных уравнений, когда  $m < 4n/3$ .

Рассмотренный нами метод Клайна [20\*] отличается от метода Питерса—Уилкинсона тем, что задача наименьших квадратов  $Ly \cong P^T b$  решается (вместо перехода к нормальным уравнениям) своеобразным вариантом метода отражений, учитывающим форму матрицы  $L$ . В [56\*] предлагается на этом этапе вместо отражений использовать модифицированный процесс Грама—Шмидта.

Главная идея метода Питерса—Уилкинсона состоит в том, чтобы заменить решение "плохо обусловленной" системы нормальных уравнений

$$A^T A x = A^T b$$

решением "хорошо обусловленной" системы

$$L^T L y = L^T P^T b. \quad (31)$$

Надежда на хорошую обусловленность системы (31) опирается на то, что все элементы матрицы  $L$  по абсолютной величине не превосходят единицу — это обеспечивается матрицей перестановок  $P$ . Практика метода свидетельствует, что эта надежда обычно оправдывается (хотя с точки зрения теории существуют треугольные матрицы, ненулевые элементы которых равны 1 или -1, с числом обусловленности порядка  $2^n$ ; такова, например, первая матрица Кахана из гл. 6). Так как, однако, число операций в методе Питерса—Уилкинсона (в главном члене) такое же, что и в методе отражений, где нет проблем с обусловленностью, то для задач с хранимыми матрицами первый метод не обладает никакими преимуществами. Современный интерес к нему связан с тем, что для больших систем разреженности обычно удается сохранить лучше при неортогональных преобразованиях. "Разреженная" трактовка метода Питерса—Уилкинсона дана в [16\*].

Глава 20. Оценки возмущения для решения линейной задачи наименьших квадратов с линейными ограничениями-равенствами получены в [27\*]. Эти оценки обобщают выведенные в гл. 8, 9 оценки возмущения для решения обычной задачи НК.

Пусть система  $Cx = d$  совместна, и пусть  $x$  — решение задачи

$$\min_F \|Ex - f\|, \quad F = \{x \mid Cx = d\}. \quad (32)$$

Составляя функцию Лагранжа

$$\Phi(x, \lambda) = \|Ex - f\|^2 + (\lambda, Cx - d),$$

где  $\lambda$  —  $m_1$ -вектор множителей Лагранжа, и приравняв нулю  $\text{grad}_x \Phi$ , получим

$$E^T(Ex - f) + C^T \lambda = 0. \quad (33)$$

Вводя вектор невязки  $r$ , отвечающий решению  $x$ :

$$r = Ex - f, \quad (34)$$

перепишем (33) в виде

$$C^T \lambda + E^T r = 0. \quad (35)$$

Вместе с уравнением связи  $Cx = d$  соотношения (34), (35) означают, что вектор

$$u = \begin{bmatrix} \lambda \\ r \\ x \end{bmatrix}$$

является решением системы

$$Bu = \begin{bmatrix} 0 & 0 & C \\ 0 & -I & E \\ C^T & E^T & 0 \end{bmatrix} u = \begin{bmatrix} d \\ f \\ 0 \end{bmatrix} = h. \quad (36)$$

Это решение единственно, если  $\ker C \cap \ker E = \{0\}$ . Заметим, что матрица  $B$  квадратная и в случае единственного решения будет не вырождена.

Будем считать, что  $C$  — матрица полного строчного ранга:  $\text{rank } C = m_1$ . Наряду с задачей (32) рассмотрим задачу

$$\min_{\tilde{F}} \|\tilde{E}x - \tilde{f}\|, \quad \tilde{F} = \{x \mid \tilde{C}x = \tilde{d}\}, \quad (37)$$

где  $\tilde{E} = E + \epsilon_E H_E$ ,  $\tilde{C} = C + \epsilon_C H_C$ ,  $\tilde{f} = f + \epsilon_f h_f$ ,  $\tilde{d} = d + \epsilon_d h_d$ ,  $\|H_E\| = \|H_C\| = \|h_f\| = \|h_d\| = 1$ ;  $\epsilon_E, \epsilon_C, \epsilon_f, \epsilon_d > 0$ . Предполагаем, что  $\text{rank } \tilde{C} = m_1$ , и обе задачи (32), (37) имеют единственное решение:  $x$  и  $\tilde{x} = x + dx$  соответственно. Построенный по  $\tilde{x}$  вектор  $\tilde{u}$  будет решением системы с матрицей  $\tilde{B}$ , аналогичной системе (36). Положим

$$H_B = \tilde{B} - B = \begin{bmatrix} 0 & 0 & \epsilon_C H_C \\ 0 & 0 & \epsilon_E H_E \\ \epsilon_C H_C^T & \epsilon_E H_E^T & 0 \end{bmatrix}.$$

Малость возмущения будет определяться требованием  $\|B^{-1}H_B\| < 1$ .

Введем обозначения:

$$\kappa_C(E) = \|E\| \|(EP_C)^*\|, \quad P_C = I - C^*C,$$

$$\kappa_E(C) = \|C\| \|C_{I,E}^*\|$$

(по поводу обозначения  $C_{I,E}^*$  см. комментарий к гл. 7),

$$\nu(E, C) = \|EC_{I,E}^*\| \frac{\|C\|}{\|E\|},$$

$$\rho_E = \frac{\|r\|}{\|E\| \|x\|},$$

$$\gamma_E = \frac{\|f\|}{\|E\| \|x\|}.$$

Оценки Элдена [27\*] имеют вид

$$\begin{aligned} \|dx\| &\leq \|(EP_C)^*\|^2 (\epsilon_C \|\lambda\| + \epsilon_E \|r\|) + \\ &+ (\epsilon_C \|C_{I,E}^*\| + \epsilon_E \|(EP_C)^*\|) \|x\| + \\ &+ \epsilon_d \|C_{I,E}^*\| + \epsilon_f \|(EP_C)^*\| + O(\epsilon^2), \end{aligned}$$

$$\begin{aligned} \frac{\|dx\|}{\|x\|} &\leq \kappa_C^2(E) \left( \nu(E, C) \frac{\epsilon_C}{\|C\|} + \frac{\epsilon_E}{\|E\|} \right) \rho_E + \\ &+ \kappa_C(E) \left( \frac{\epsilon_E}{\|E\|} + \frac{\epsilon_f}{\|f\|} \gamma_E \right) + \kappa_E(C) \left( \frac{\epsilon_C}{\|C\|} + \frac{\epsilon_d}{\|d\|} \right) + O(\epsilon^2). \end{aligned}$$

Коэффициенты  $\kappa_C(E)$ ,  $\kappa_E(C)$  Элден называет числами обусловленности задачи (32). Он отмечает, что эти числа могут быть невелики даже при плохо обусловленной  $E$ . Тем самым задача (32) может оказаться хорошо обусловленной, хотя соответствующая задача без ограничений  $\min_x \|Ex - f\|$  обусловлена плохо (разумеется, может быть и наоборот).

Поэтому вряд ли стоит начинать решение задачи (32) с преобразования матрицы  $E$ . Тот же вывод, впрочем, содержится в более ранней работе [43\*].

Анализ чувствительности решения задачи (32) к возмущениям исходных данных проводится и в [57\*]; при этом применяются только обычные псевдообратные матрицы.

Представление (36) используется в [17\*] для построения алгоритма итерационного уточнения решения задачи (32). Полагая  $u^{(0)} = 0$ , можем описать процесс формулами ( $s = 0, 1, \dots$ ):

$$a) v^{(s)} = h - Bu^{(s)};$$

$$б) \delta u^{(s)} = B^{-1}v^{(s)};$$

$$в) u^{(s+1)} = u^{(s)} + \delta u^{(s)}.$$

При  $s > 0$  в векторе

$$v^{(s)} = \begin{bmatrix} d^{(s)} \\ f^{(s)} \\ g^{(s)} \end{bmatrix}$$

подвектор  $g^{(s)}$  не будет нулевым. Заменяя в векторе  $h$  нулевой подвектор вектором  $g$ , укажем, следуя [17], метод решения системы (36). При этом считаем, что  $\text{rank } C = m_1$  и  $\ker C \cap \ker E = \{0\}$ .

Предположим (переставляя в случае необходимости столбцы  $C$ ), что в представлении  $C = [C_{11} : C_{12}]$  подматрица  $C_{11}$  квадратная и невырожденная. Пусть  $Q_{11}$  — ортогональная матрица порядка  $m_1$  такая, что в матрице

$$Q_{11}C = [R_{11} : R_{12}] \quad (38)$$

$R_{11}$  верхняя треугольная. Разбивая  $E$  аналогично  $C$ :

$$E = \begin{bmatrix} E_{11} & E_{12} \\ m_1 & n-m_1 \end{bmatrix},$$

положим

$$Q_{12} = (R_{11}^T)^{-1} E_{11}^T, \quad \tilde{E}_{12} = E_{12} - Q_{12}^T R_{12}. \quad (39)$$

Введем еще матрицу  $Q_{22}$  порядка  $m_2$  — ортогональный сомножитель  $QR$ -разложения  $\tilde{E}_{12}$ :

$$Q_{22} \tilde{E}_{12} = \begin{bmatrix} R_{22} \\ 0 \end{bmatrix} \begin{matrix} n-m_1 \\ m-n \end{matrix}, \quad m = m_1 + m_2. \quad (40)$$

Матрицу размера  $(n-m_1) \times m_2$ , образованную верхними  $n-m_1$  строками  $Q_{22}$ , обозначим через  $\tilde{Q}_{22}$ . Полагая

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

запишем соотношения (38)–(40) в форме матричного равенства

$$\begin{bmatrix} C \\ E \end{bmatrix} = \begin{bmatrix} Q_{11}^T & 0 \\ Q_{12}^T & \tilde{Q}_{22}^T \end{bmatrix} R. \quad (41)$$

Векторы  $x, g$  и последующие векторы той же размерности будем представлять в виде

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \begin{matrix} m_1 \\ n-m_1 \end{matrix}, \quad g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \begin{matrix} m_1 \\ n-m_1 \end{matrix}.$$

Умножая нижнюю блочную строку системы (36) слева на  $R$ , полагая  $Rg = t$  и выполняя замену переменных  $y = Rx$ , получим (см. (41))

$$\begin{bmatrix} 0 & 0 & Q_{11}^T & 0 \\ 0 & -I & Q_{12}^T & \tilde{Q}_{22}^T \\ Q_{11} & Q_{12} & 0 & 0 \\ 0 & \tilde{Q}_{22} & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ r \\ y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} d \\ f \\ t_1 \\ t_2 \end{bmatrix} \quad (42)$$

Пользуясь ортогональностью  $Q_{11}$ , находим  $y_1$  из верхней блочной

строки (42):

$$y_{11} = Q_{11} t_1.$$

Построим вспомогательный  $m_2$ -вектор  $s$ :

$$s = -Q_{22} (f - Q_{12}^T y_1) = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \begin{matrix} n-m_1 \\ m-n \end{matrix}.$$

Последняя блочная строка (42)  $\tilde{Q}_{22} r = t_2$  показывает, что верхние  $n-m_1$  компонент вектора  $Q_{22} r$  образуют вектор  $t_2$ . Из второй же блочной строки  $-r + Q_{12}^T y_1 + \tilde{Q}_{22}^T y_2 = f$  следует

$$Q_{22} r = s + Q_{22} \tilde{Q}_{22}^T y_2. \quad (43)$$

Поскольку произведение  $Q_{22} \tilde{Q}_{22}^T$  имеет вид

$$Q_{22} \tilde{Q}_{22}^T = \begin{bmatrix} I_{n-m_1} \\ 0 \end{bmatrix},$$

равенство (43) означает, что последние  $m-n$  компонент  $Q_{22} r$  составляют

вектор  $s_2$ . Итак,  $Q_{22} r = \begin{bmatrix} t_2 \\ s_2 \end{bmatrix}$ , откуда в силу ортогональности  $Q_{22}$

$r = Q_{22}^T \begin{bmatrix} t_2 \\ s_2 \end{bmatrix}$ . То же равенство (43) позволяет определить  $y_2$ :  $y_2 = t_2 - s_1$ .

Наконец,  $\lambda$  находим из третьей блочной строки (42)

$$\lambda = Q_{11}^T (t_1 - Q_{12} r).$$

Алгоритмическая процедура описанного процесса для случая, когда необходимые ортогонально-треугольные разложения выполняются методом отражений, приведена в [17\*]. Во второй части работы [14\*] дана программа этого же процесса, в которой для разложений используется модифицированный метод Грама–Шмидта.

Глава 22. Введенный в комментарии к гл. 7 аппарат взвешенных псевдообратных матриц очень облегчает доказательство сходимости метода взвешивания [28]. Прежде всего, если система  $Cx = d$  совместна, то решение минимальной длины задачи (32) выражается формулой

$$x = C_{I,E}^* d + (EP_C)^* f.$$

Здесь  $P_C$  — проектор на ядро  $C$ :  $P_C = I - C^* C$ . Вектор  $x$  тогда и только тогда будет единственным решением (32), когда  $\ker C \cap \ker E = \{0\}$ . Будем считать это условие выполненным.

Решение  $x_e$  задачи

$$\begin{bmatrix} C \\ \epsilon E \end{bmatrix} x \cong \begin{bmatrix} d \\ \epsilon f \end{bmatrix}$$

можно определить из нормальной системы

$$(C^T C + \epsilon^2 E^T E) x = C^T d + \epsilon^2 E^T f,$$

т. е.

$$x_e = (C^T C + \epsilon^2 E^T E)^{-1} (C^T d + \epsilon^2 E^T f).$$

Применяя к первому слагаемому правой части формулу (13), а ко второму формулу (14), получим

$$\lim_{\epsilon \rightarrow 0} x_e = C_{I,E}^* d + (EP_C)^* f = x.$$

Вычислительные трудности, возникающие при реализации метода взвешивания, анализируются (посредством обобщенного сингулярного разложения) в [65\*]. Там же описаны два приема — экстраполяция и итерационное уточнение, позволяющие во многих случаях получить приемлемое приближение к решению задачи (32) при не слишком малых значениях весового параметра  $\epsilon$ .

Глава 23. Для задачи

$$\min_F \|Ex - f\|, \quad F = \{x \mid Gx \geq h\} \quad (44)$$

анализ чувствительности решения к возмущениям входных данных проводится в [44\*]. Численным методам для задачи (44) посвящены работы [58\*, 45\*]. Во второй из них рассматривается случай ограничений специального вида:  $c_i \leq x_i \leq d_i$ ,  $i = 1, \dots, k$ ;  $x_i \geq c_i$ ,  $i = k+1, \dots, l$ .

Главы 24, 27. Устойчивость методов перестройки  $QR$ -разложения, предполагающих хранение ортогонального сомножителя, анализируется в [54\*]. Вот итоги этого анализа.

Как отмечено в гл. 15 (см., в частности, формулы (15.39), (15.40)), матрица  $\bar{A}_{k+1}$ , полученная в результате применения к  $m \times n$ -матрице  $A$  последовательности  $k$  левых преобразований Хаусхолдера, может быть представлена в виде

$$\begin{aligned} \bar{A}_{k+1} &= Q_k \dots Q_1 (A + H_k) = Q(A + H_k), \\ Q &= Q_k \dots Q_1, \quad \bar{A}_1 = A. \end{aligned} \quad (45)$$

Таким образом,  $\bar{A}_{k+1}$  можно рассматривать как продукт *точного* ортогонального преобразования возмущенной матрицы  $A + H_k$ . Для нормы матрицы эквивалентного возмущения  $H_k$  выполняется оценка

$$\|H_k\|_F \leq O(\eta) \|A\|_F. \quad (46)$$

Вид коэффициента при  $\eta$ , приведенный в (15.40), для нас сейчас не важен. Такого же типа оценки справедливы и для других видов ортогональных

преобразований — последовательностей левых, правых или двусторонних отражений либо вращений.

Алгоритмы, дающие результат, являющийся точным преобразованием слабо возмущенной исходной матрицы, в вычислительной линейной алгебре принято считать устойчивыми. Как показано в [54\*], методы перестройки  $QR$ -разложения из гл. 24 устойчивы именно в этом смысле. Некоторую особенность составляет случай удаления строки.

Для простоты будем считать, что  $m > n$  и  $\text{rank } A = n$ . Пусть  $Q$  и  $R$  дают приближенное ортогонально-треугольное разложение матрицы  $A$ , т. е.

$$P(A + H) = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad \|H\|_F \leq O(\eta) \|A\|_F, \quad \|Q - P\| \leq O(\eta),$$

где  $P$  — ортогональная матрица, а  $R$  — невырожденная верхняя треугольная матрица. Таким образом, матрица  $Q$  не предполагается точно ортогональной, но должна быть близка к некоторой ортогональной матрице  $P$ .

Пусть  $\tilde{A}$  — матрица, полученная из  $A$  удалением какой-либо строки. Первый метод модификации из § 5 гл. 27 приводит к построению почти-ортогональной матрицы  $\tilde{Q}$  порядка  $m-1$  и новой треугольной  $n \times n$ -матрицы  $\tilde{R}$ . Почти-ортогональность  $\tilde{Q}$  означает существование ортогональной матрицы  $\tilde{P}$  такой, что  $\|\tilde{Q} - \tilde{P}\| \leq O(\eta)$ ; при этом можно показать, что

$$\tilde{P}(\tilde{A} + \tilde{H}) = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix}, \quad \|\tilde{H}\|_F \leq O(\eta) \|A\|_F.$$

Обратим внимание, что норма эквивалентного возмущения оценивается через норму матрицы  $A$ , а не  $\tilde{A}$ . Если бы к  $\tilde{A}$  был заново применен алгоритм  $QR$ -разложения (а не алгоритм перестройки  $QR$ -разложения  $A$ ), то, согласно (45), (46), в правой части оценки была бы норма самой матрицы  $\tilde{A}$ . Указанное различие может оказаться существенным, если из  $A$  удаляется строка с наибольшей нормой, так что уровень элементов в  $\tilde{A}$  становится значительно ниже, чем в  $A$ .

В [63\*] дан обзор методов вычисления ортогонально-треугольного разложения и его перестройки при модификациях ранга 1 для задач с матрицами ленточной структуры.

Глава 25. Способ сведения взвешенных наименьших квадратов к обычной задаче НК, описанный в конце § 2, может оказаться несостоятельным во (вполне реалистической) ситуации, когда ковариационная матрица  $C$  вырождена. Пэйджем [52\*] указана эквивалентная формулировка, сохраняющая смысл и при вырожденной  $C$ ; на основе этой формулировки им построен устойчивый численный алгоритм.

Пусть вначале матрица  $C$  не вырождена и

$$C = FF^T \quad (47)$$

есть ее разложение Холесского. Задача

$$\min_{x \in R^n} \|F^{-1}(Ax - b)\|$$

эквивалентна задаче

$$\min_v \|v\|, \quad b = Ax + Fv. \quad (48)$$

Если  $C$  вырождена, но сохраняет положительную полуопределенность, то ее все еще можно представить в виде (47), где теперь  $F$  есть  $m \times k$ -матрица,  $k = \text{rank } C$ . При этом постановка (48) по-прежнему осмыслена, если  $v$  считать  $k$ -мерным вектором, а систему  $b = Ax + Fv$  — совместной.

Ограничимся ради простоты случаем, когда матрица  $A$  имеет полный столбцовый ранг. Матрицу  $Q$  из ортогонально-треугольного разложения  $A$

$$A = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \quad (49)$$

разобьем на подматрицы  $Q_1, Q_2$ :

$$Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}, \quad \begin{matrix} n & m-n \end{matrix}$$

Умножая обе части равенства  $Ax + Fv = b$  на  $Q^T$ , получим

$$R_1 x + Q_1^T F v = Q_1^T b, \quad (50)$$

$$Q_2^T F v = Q_2^T b. \quad (51)$$

Если  $v$  известно, то (50) превращается в треугольную систему, однозначно определяющую  $x$ . Согласно (48), наша задача свелась к вычислению нормального решения (= решения минимальной длины) для совместной системы (51).

Полагая  $v = Pu$ , где  $P$  — ортогональная матрица, выберем  $P$  так, чтобы матрица  $Q_2^T F P$  приобрела вид

$$Q_2^T F P = \begin{bmatrix} 0 & S \end{bmatrix}. \quad (52)$$

Матрица  $S$  в (52) должна иметь полный столбцовый ранг, равный  $k$ . В соответствии с (52) разобьем матрицу  $P$  и вектор  $u$ :

$$P = \begin{bmatrix} P_1 & P_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \begin{matrix} k-l & l \end{matrix}$$

Система (51) перейдет в

$$S u_2 = Q_2^T b. \quad (53)$$

Если рассматривать (53) как систему относительно  $u_2$ , то ее решение  $\hat{u}_2$

определено однозначно, поскольку  $S$  — матрица полного столбцового ранга. Вектор  $\hat{u}_2$  может быть вычислен посредством алгоритма HFT — HS1 из гл. 11. Та же система (53), рассматриваемая на этот раз относительно вектора  $u$ , имеет нормальное решение

$$\hat{u} = \begin{bmatrix} 0 \\ \hat{u}_2 \end{bmatrix},$$

откуда следует, что нормальным решением (51) будет вектор

$$\hat{v} = P_2 \hat{u}_2. \quad (54)$$

Подставляя (54) в (50), найдем решение  $x$  исходной задачи.

В [52\*] проведен анализ чувствительности решения задачи (48) к малым возмущениям ее коэффициентов, а также анализ численной устойчивости изложенного выше алгоритма. Вот основные выводы этой работы.

Будем считать, что возмущения коэффициентов настолько малы, что не изменяют ни ранга матрицы  $A$ , ни ранга  $l$  матрицы  $S$  в (52). В наихудшем случае погрешность в решении задачи (48) пропорциональна произведению погрешностей в коэффициентах и числа  $\sigma_{\min}^{-2}(A) \sigma_{\min}^{-1}(S)$ ;  $\sigma_{\min}(\cdot)$  обозначает минимальное сингулярное число соответствующей матрицы. Полезно заметить, что  $\sigma_{\min}(S) = \sigma_{\min}(Q_2^T F)$ ; при этом, как следует из  $QR$ -разложения  $A$ , столбцы  $Q_2$  образуют (ортогональный) базис подпространства  $(\text{Im } A)^\perp = \ker A^T$ . Таким образом, произведение  $Q_2^T F$  указывает, в частности, в какой степени неортогональны к образу  $A$  столбцы весовой матрицы  $F$ . В случае обычной задачи НК  $F = I_m$ ,  $\sigma_{\min}(Q_2^T) = 1$ , и мы приходим к известному из гл. 9 факту появления в оценке для  $dx$  квадратичного по  $\|A^*\|$  члена.

В некоторых ситуациях оценка для  $dx$  более оптимистична. Если, например,  $Q_2^T \delta A = 0$ , т.е. возмущения в коэффициентах матрицы  $A$  не изменяют ее образа, то обусловленность задачи определяется произведением  $[\sigma_{\min}(A) \sigma_{\min}(S)]^{-1}$ . Если же все возмущения  $\delta A$ ,  $\delta F$ ,  $\delta b$  ортогональны к  $Q_2$ , то коэффициентом пропорциональности будет  $\sigma_{\min}^{-1}(A)$ .

Анализ алгоритма (49)–(54) проделан в [52\*] при дополнительном предположении, что  $F$  — квадратная и, следовательно, невырожденная матрица. Для этого случая показано, что вычисленное решение  $\tilde{x}$  будет точным для задачи, отличающейся от (48) малым возмущением коэффициентов. Таким образом, алгоритм устойчив в смысле обратного анализа погрешностей.

Пэйдж отмечает, что, поскольку условие  $b = Ax + Fv$  должно удовлетворяться точно (а не в смысле наименьших квадратов!), для упрощения внешнего вида матриц  $A$ ,  $F$  необязательно использовать ортогональные преобразования.

В [53\*] описаны алгоритмы решения задачи (48) (в том числе включающие неортогональные преобразования) для нескольких ситуаций, когда



матрица  $F$  имеет какую-либо специальную форму, например является невырожденной нижней треугольной либо блочно диагональной матрицей. Алгоритм, построенный для первого случая, требует  $\approx (m^2 n/2 + mn^2 - 2n^3/3)p$  операций умножения. Если используются обычные вращения, то  $p$  в этом выражении нужно положить равным 4. Неучет треугольной формы матрицы  $F$  приводит к тому, что в формуле для числа операций возникает слагаемое порядка  $m^3$ .

В [51\*] указаны численно устойчивые и эффективные алгоритмы перестройки решения задачи (48) при удалении столбца из матрицы или, наоборот, введении нового столбца. В [42\*] анализируются статистические свойства формулировки (48) и устанавливается связь этого подхода с предложенным Рао методом обобщенного обращения фундаментальной матрицы

$$\begin{bmatrix} C & A \\ A^T & 0 \end{bmatrix}$$

Обзор методов численного решения задачи (48) дан в [36].

Упомянутая в конце § 4 задача о выборе  $\lambda$  из условия, чтобы норма решения либо норма невязки имела заранее предписанное значение, в более общей постановке рассматривается в [31]: для заданных  $m \times n$ -матрицы  $A$ ,  $p \times n$ -матрицы  $C$ ,  $m$ -вектора  $b$ ,  $p$ -вектора  $d$  и положительного числа  $\alpha$  найти вектор  $x$ , минимизирующий  $\|Ax - b\|$  при условии  $\|Cx - d\| = \alpha$ . Будем называть эту задачу задачей наименьших квадратов с квадратичным ограничением (сокращенно ЗНККО). Предполагается, что множество  $\{x \mid \|Cx - d\| = \alpha\}$  непусто и

$$\alpha > \min_{x \in R^n} \|Cx - d\|.$$

Это условие обеспечивает разрешимость ЗНККО. Чтобы число решений было конечным, потребуем, чтобы  $\ker A \cap \ker C = \{0\}$ .

Метод Лагранжа, примененный к ЗНККО, дает следующие нормальные уравнения:

$$(A^T A + \lambda C^T C)x = A^T b + \lambda C^T d,$$

$$\|Cx - d\|^2 = \alpha^2.$$

Среди решений  $(\lambda, x(\lambda))$  системы нормальных уравнений решение ЗНККО дает пара  $(\lambda_0, x(\lambda_0))$  с наибольшим значением  $\lambda$ . Если, в частности, имеется пара с положительным  $\lambda$ , то такая пара будет ровно одна, и соответствующий единственный вектор  $x(\lambda)$  решает ЗНККО. Если же все  $\lambda$  отрицательны, то решение ЗНККО может оказаться неединственным. Так будет, если  $\lambda_0$  — собственное значение пучка  $A^T A + \lambda C^T C$ .

Интересно отметить, что исследование системы нормальных уравнений проводится в [31] посредством обобщенного сингулярного разложения пары  $(A, C)$ .

Частными случаями ЗНККО являются задачи:

$$\text{минимизировать } \|Ax - b\|$$

(55)

$$\text{при условии } \|x\| = \alpha$$

(эта задача рассматривалась еще в [30]) и

$$\text{минимизировать } \|x\|$$

(56)

$$\text{при условии } \|Cx - d\| = \alpha.$$

Последняя интерпретируется геометрически как определение точки гиперэллипсоида, ближайшей к началу координат.

К ЗНККО тесно примыкает задача

$$\text{минимизировать}$$

$$\|Ax - b\|$$

$$\text{при условии } \|Cx - d\| \leq \alpha.$$

(57)

В самом деле, либо  $\|C\tilde{x} - d\| < \alpha$  для некоторого псевдорешения  $\tilde{x}$  задачи  $Ax \cong b$ , и в этом случае одним из решений (57) будет  $\tilde{x}$ , либо минимум достигается на границе, и мы приходим к соответствующей ЗНККО.

В [28] рассмотрены задачи, получающиеся из (55) либо (56) заменой ограничения-уравнения ограничением-неравенством и заменой евклидовой длины в целевой функции либо в левой части ограничения на эллипсоидальную полуорму.

Некоторый метод выбора гребневого параметра  $\lambda$  в задаче (25.31) предложен в § 12.1 книги [34]. Там же на стр. 418 описан следующий алгоритм выбора подмножества (см. § 5):

1. Вычисляется сингулярное разложение матрицы  $A$ :

$$U^T A V = \text{diag}(\sigma_1, \dots, \sigma_n).$$

Матрица  $V$  должна быть сохранена.

2. Определяется значение  $k$  псевдоранга, после чего  $V$  разбивается в соответствии с  $k$ :

$$V = \left[ \begin{array}{cc} V_{11} & V_{12} \\ \underbrace{V_{21}}_k & \underbrace{V_{22}}_{n-k} \end{array} \right] \begin{array}{l} k \\ n-k \end{array}$$

3. К матрице  $[V_{11}^T : V_{21}^T]$  применяется алгоритм HFTI из гл. 14:

$$Q[V_{11}^T : V_{21}^T]P = [R_{11} : R_{12}];$$

здесь  $P$  — матрица, описывающая произведенные перестановки столбцов.

4. Положим  $AP = \left[ \begin{array}{cc} B_1 & B_2 \\ \underbrace{\phantom{B_1}}_k & \underbrace{\phantom{B_2}}_{n-k} \end{array} \right]$ . Решается задача наименьших квадратов

$$B_1 z \cong b.$$

(58)

Таким образом, мощность выбираемого подмножества определяется на этапе 2, а само это подмножество столбцов указывает матрица  $P$ . На этапе 4 решается редуцированная задача (58).

Насколько велика невязка задачи (58) по сравнению с невязкой исходной задачи  $Ax \cong b$ ? Если  $\hat{z}$ ,  $\hat{x}$  — решения обеих задач,  $r_z$ ,  $r_x$  — отвечающие им невязки, то

$$\|r_z - r_x\| \leq \frac{\sigma_{k+1}}{\sigma_k} \|R_{11}^{-1}\| \|b\|. \quad (59)$$

Сингулярные числа  $\sigma_1, \dots, \sigma_n$  предполагаются упорядоченными по убыванию. Понятно, что  $\|r_z\| \geq \|r_x\|$ .

Оценка (59) объясняет, почему для выбора подмножества был применен алгоритм HFTI: он позволяет определить в достаточной степени невырожденную подматрицу порядка  $k$  в первых  $k$  столбцах  $V$ , т.е. по возможности уменьшить значение  $\|R_{11}^{-1}\|$ . Есть и другая причина: необходимость обеспечить линейную независимость столбцов  $B_1$ . Справедливы неравенства

$$\frac{\sigma_k}{\|R_{11}^{-1}\|} < \sigma_{\min}(B_1) < \sigma_k.$$

Чем меньше  $\|R_{11}^{-1}\|$ , тем больше "запасом" линейной независимости обладают столбцы  $B_1$ .

## СПИСОК ЛИТЕРАТУРЫ

1. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. — М.: Наука, 1977.
2. Гордонова В.И. Оценки ошибок округления при решении системы уравнений. — ЖВМ и МФ, 1969, 9, № 4, с. 775–782.
3. Гэйл Д. Теория линейных экономических моделей. — М.: ИЛ, 1963.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. — М.: Статистика, 1973.
5. Ортега Дж., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.
6. Тихонов А.Н., Гласко В.Б. О приближенном решении интегральных уравнений Фредгольма 1 рода. — ЖВМ и МФ, 1964, 4, № 3, с. 564–574.
7. Уилкинсон Дж. Алгебраическая проблема собственных значений. — М.: Наука, 1970.
8. Уилкинсон Дж., Райнш К. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. — М.: Машиностроение, 1976.
9. Фаддеев Д.К., Кублановская В.Н., Фаддеева В.Н. О решении линейных алгебраических систем с прямоугольными матрицами. — Тр. Мат. ин-та АН СССР, 1968, 96, с. 76–92.
10. Фиакко А., Мак-Кормик Дж. Нелинейное программирование: методы последовательной безусловной минимизации. — М.: Мир, 1972.
11. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.
12. Халмош П.Р. Конечномерные векторные пространства. — М.: Физматгиз, 1963.
13. ANSI Subcommittee X3J3. Clarification of Fortran standard—second report. — Commun. ACM, 1971, 14, № 10, p. 628–642.
14. ASA Committee X3. Fortran vs. Basic Fortran. — Commun. ACM., 1964, 7, № 10, p. 591–625.
15. Autonne L. Sur les matrices hypohermitiennes et sur les matrices unitaires. — Ann. Univ. Lyon, 1915, 38, p. 1–77.
16. Bard Y. Comparison of gradient methods for the solution of nonlinear parameter estimation problems. — SIAM J. Numer. Anal., 1970, 7, № 1, p. 157–186.
17. Bartels R.H. A stabilization of the simplex method. — Numer. Math., 1971, 16, p. 414–434.
18. Bartels R.H., Golub G.H., Saunders M.A. Numerical techniques in mathematical programming. — Nonlinear programming. — New York: Academic Press, 1970, p. 123–176.
19. Bauer F.L. Optimal scaling of matrices and the importance of the minimal condition. — Proc. IFIP Congr. 1962. — Amsterdam: North Holland, 1963, p. 198–201.
20. Ben-Israel A. A modified Newton–Raphson method for the solution of systems of equations. — Israel J. Math., 1965, 3, p. 94–98.
21. Ben-Israel A. On error bounds for generalized inverses. — SIAM J. Numer. Anal., 1966, 3, № 4, p. 585–592.
22. Bennett J.M. Triangular factors of modified matrices. — Numer. Math., 1965, 7, p. 217–221.

23. Björck A. Solving linear least squares problems by Gram-Schmidt orthogonalization. — BIT, 1967, 7, p. 1–21.
24. Björck A. Iterative refinement of linear least squares solutions I. — BIT, 1967, 7, p. 257–278.
25. Björck A., Golub G.H. Iterative refinement of linear least squares solutions by Householder transformation. — BIT, 1967, 7, p. 322–337.
26. Björck A., Golub G.H. Numerical methods for computing angles between linear subspaces. — Math. Comp., 1973, 27, p. 579–594.
27. Boggs D.H. A partial-step algorithm for the non-linear estimation problem. — AIAA J., 1972, 10, № 5, p. 675–679.
28. Bouffion T., Odell P. Generalized inverse matrices. — New York: Wiley, 1971.
29. Brown K.M. Computer solution of nonlinear algebraic equations and nonlinear optimization problems. — Proc. Share, XXXVII. — New York, 12 p.
30. Buchanan J.E., Thomas D.H. On least-squares fitting of twodimensional data with a special structure. — SIAM J. Numer. Anal., 1968, 5, № 2, p. 252–257.
31. Businger P.A. Matrix scaling with respect to the maximum norm, the sum-norm, and the euclidean norm. — Texas Univ., Austin. Thesis TNN71, 1967, 119 p.
32. Businger P.A. MIDAS-solution of linear algebraic equations. — Bell Teleph. Lab., Numer. Math. Comp. Programs 3, issue 1. — Murray Hill, N.J., 1970.
33. Businger P.A. Updating a singular value decomposition. — BIT, 1970, 10, p. 376–385.
34. Businger P.A., Golub G.H. An Algol procedure for computing the singular value decomposition. — Stanford Univ. Rept. № CS-73. — Calif., Stanford, 1967.
35. Businger P.A., Golub G.H. Singular value decomposition of a complex matrix. — Commun. ACM, 1969, 12, № 10, p. 564–565.
36. Carasso C., Laurent P.J. On the numerical construction and the practical use of interpolating spline-functions. — Proc. IFIP Congr., 1968. — Amsterdam: North Holland, 1969.
37. Chambers J.M. Regression updating. — J. Amer. Stat. Ass., 1971, 66, p. 744–748.
38. Cheney E.M., Goldstein A.A. Mean-square approximation by generalized rational functions. — Boeing Sci. Res. Lab. Math. Note, № 465. — Seattle, 1966, p. 17.
39. Cody W.J. Software for the elementary functions. — Mathematical Software. — New York: Academic Press, 1971, p. 171–186.
40. Cowden D.J. A procedure for computing regression coefficients. — J. Amer. Stat. Ass., 1958, 53, p. 144–150.
41. Cox M.G. The numerical evaluation of  $B$ -splines. — Nath. Phys. Lab. Rept. № NPL-DNAC-4. — England, Middlesex, Teddington, 1971, 34 p.
42. Cruise D.R. Optimal regression models for multivariate analysis (factor analysis). — Naval Weapons Center Rept. NWCTP 5103. — Calif., China Lake, 1971, 60 p.
43. Davidson W.C. Variance algorithm for minimization. — Optimization. — London: Academic Press, 1968, p. 13–20.
44. Davis Ch.D., Kahan W.M. The rotation of eigenvectors by a perturbation III. — SIAM J. Numer. Anal., 1970, 7, № 1, p. 1–46.
45. De Boor C. Subroutine package for calculating with  $B$ -splines. — Los Alamos Sci. Lab. Rept. № LA-4728-MS, 1971, 12 p.
46. De Boor C. On calculating with  $B$ -splines. — J. Approximation Theory, 1972, 6, № 1, p. 50–62.
47. De Boor C., Rice J.R. Least squares cubic spline approximation I — fixed knots. — Purdue Univ. Rept. CSD TR 20. Ind., Lafayette, 1968, 30 p.
48. De Boor C., Rice J.R. Least squares cubic spline approximation II — variable knots. — Purdue Univ. Rept. CSD TR 21. Ind., Lafayette, 1968, 28 p.
49. Dyer P., McReynolds S.R. The extension of square-root filtering to include process noise. — J. Opt.: Theory and Applications, 1969, 3, p. 92–105.
50. Eberlein P.J. Solution to the complex eigenproblem by a norm reducing Jacobi type method. — Numer. Math., 1970, 14, p. 232–245.
51. Eckart C., Young G. The approximation of one matrix by another of lower rank. — Psychometrika, 1935, 1, p. 211–218.
52. Fletcher R.H. Generalized inverse methods for the best least squares solution of systems of non-linear equations. — Comput. J., 1968, 10, p. 392–399.
53. Fletcher R. A modified Marquardt subroutine for non-linear least squares. — Amonic Energy Res. Estab. Rept. № R-6799. — England, Berkshire, Harwell, 1971, 24 p.
54. Fletcher R., Lill S.A. A class of methods for nonlinear programming, II, computational experience. — Nonlinear programming. — New York: Academic Press, 1970, p. 67–92.
55. Forsythe G.E. Generation and use of orthogonal polynomial for data-fitting with a digital computer. — J. Soc. Indust. Appl. Math., 1957, 5, p. 74–88.
56. Forsythe G.E. Today's methods of linear algebra. — SIAM Rev., 1967, 9, p. 489–515.
57. Forsythe G.E. Pitfalls in computation, or why a math. book isn't enough. Amer. Math. Monthly, 1970, 77, p. 931–956.
58. Fox L. An introduction to numerical linear algebra. — New York: Oxford Univ. Press, 1965.
59. Francis J.G.F. The QR transformation; Parts I, II. — Comput. J., 1961, 4, p. 265–271, 332–345.
60. Franklin J.N. Matrix theory. — Prentice-Hall, 1968.
61. Franklin J.N. Well-posed stochastic extensions of ill-posed linear problems. — J. Math. Anal. Appl., 1970, 31, p. 682–716.
62. Gale D. How to solve least inequalities. — Amer. Math. Monthly, 1969, 76, p. 589–599.
63. Gar side M.J. Some computational procedures for the best subset problem. — Appl. Stat., 1971, 20, p. 8–15, 111–115.
64. Gastinel N. Linear numerical analysis. — New York: Academic Press, 1971.
65. Gentleman W.M. Least squares computations by Givens transformations without square roots. — Univ. of Waterloo Rept. CSRR-2062. — Canada, Ontario, Waterloo, 1972, 17 p.
66. Gentleman W.M. Basic procedures for large, sparse or weighted linear least squares problems. — Univ. of Waterloo Rept. CSRR-2068. — Canada, Ontario, Waterloo, 1972, 14 p.
67. Gill P.E., Murray W. A numerically stable form of the simplex algorithm. — Nat. Phys. Lab. Math. Tech. Rept. № 87. — England, Middlesex, Teddington, 1970, 43 p.
68. Gill P.E., Golub G.H., Murray W., Saunders M.A. Methods for modifying matrix factorizations. — Stanford Univ. Rept. № CS-322. — Calif., Stanford, 1972, 60 p.
69. Givens W. Numerical computation of the characteristic values of a real symmetric matrix. — Oak Ridge Nat. Lab. Rept. ORNL-1574. — Tenn., Oak Ridge, 1954, 107 p.
70. Golub G.H. Least squares, singular values and matrix approximations. — Aplikace Matematiky, 1968, 13, p. 44–51.
71. Golub G.H. Matrix decompositions and statistical calculations. — In: Statistical Calculations. — New York: Academic Press, 1969, p. 365–397.
72. Golub G.H., Businger P.A. Linear least squares solutions by Householder transformations. — Numer. Math., 1965, 7, p. 269–276.
73. Golub G.H., Kahan W. Calculating the singular values and pseudoinverse of a matrix. — SIAM J. Numer. Anal., 1965, 2, № 3, p. 205–224.
74. Golub G.H., Pereyra V. The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. — SIAM J. Numer. Anal., 1973, 10, p. 413–432.
75. Golub G.H., Reinsch C. Singular value decomposition and least squares solutions. — Numer. Math., 1970, 14, № 5, p. 403–420.
76. Golub G.H., Saunders M.A. Linear least squares and quadratic programming. — Integer and nonlinear programming, II. — Amsterdam: North Holland, 1970, p. 229–256.
77. Golub G.H., Smith L.B. Chebyshev approximation of continuous functions by a Chebyshev system of functions. — Commun. ACM, 1971, 14, № 11, p. 737–746.
78. Golub G.H., Styan G.P.H. Numerical computations for univariate linear models. — J. Stat. Comp. and Simulation, 1973, 2, p. 253–274.
79. Golub G.H., Underwood R. Stationary values of the ratio of quadratic forms, subject to linear constraints. — Z. Angew. Math. Phys., 1970, 21, p. 318–326.

80. Golub G.H., Wilkinson J.H. Note on the iterative refinement of least squares solution. — Numer. Math., 1966, 9, p. 139–148.
81. Graybill F.A., Meyer C.D., Painter R.J. Note on the computation of the generalized inverse of a matrix. — SIAM Rev., 1966, 8, p. 522–524.
82. Greville T.N.E. The pseudo-inverse of a rectangular or singular matrix and its application to the solution of systems of linear equations. — SIAM Rev., 1959, 1, p. 38–43.
83. Greville T.N.E. Some applications of the pseudoinverse of a matrix. — SIAM Rev., 1960, 2, p. 15–22.
84. Greville T.N.E. Note on the generalized inverse of a matrix product. — SIAM Rev., 1966, 8, p. 518–521.
85. Halmos P.R. Introduction to Hilbert space. — New York: Chelsea Publ. Co., 1957.
86. Hanson R.J. Computing quadratic programming problems: linear inequality and equality constraints. — JPL Sec. 314 Tech. Mem. № 240. — Calif., Pasadena, California Inst. of Technology, Jet Propulsion Lab., 1970.
87. Hanson R.J. A numerical method for solving Fredholm integral equations of the first kind, using singular values. — SIAM J. Numer. Anal., 1971, 8, № 3, p. 616–622.
88. Hanson R.J. Integral equations of immunology. — Commun. ACM, 1972, 15, № 10, p. 883–890.
89. Hanson R.J., Dyer P. A computational algorithm for sequential estimation. — Comput. J., 1971, 14, № 3, p. 285–290.
90. Hanson R.J., Lawson C.L. Extensions and applications of the Householder algorithm for solving linear least squares problems. — Math. Comp., 1969, 23, № 108, p. 787–812.
91. Healy M.J.R. Triangular decomposition of a symmetric matrix. — Appl. Stat., 1968, 17, p. 195–197.
92. Hemmerle W.J. Statistical computations on a digital computer. — New York: Blaisdell Publ. Co., 1967.
93. Hestenes M.R. Inversion of matrices by biorthogonalization and related results. — J. Soc. Indust. Appl. Math., 1958, 6, p. 51–90.
94. Hestenes M.R., Stiefel E. Methods of conjugate gradients for solving linear systems. — Nat. Bur. Stand. J. Res., 1952, 49, p. 409–436.
95. Hilsenrath J., Ziegler G.G. et al. OMNITAB, a computer program for statistical and numerical analysis. — Nat. Bur. Stand., Handbook 101, 1966 (Revised 1968), 275 p.
96. Hoerl A.E. Optimum solution of many variable equations. — Chem. Eng. Progress, 1959, 55, № 11, p. 69–78.
97. Hoerl A.E. Application of ridge analysis to regression problems. — Chem. Eng. Progress, 1962, 58, № 3, p. 54–59.
98. Hoerl A.E. Ridge analysis. — Chem. Eng. Progress, 1964, 60, № 50, p. 67–78.
99. Hoerl A.E., Kennard R.W. Ridge regression: biased estimation for nonorthogonal problems. — Technometrics, 1970, 12, p. 55–67, 69–82.
100. Hoffman A.J., Wielandt H.W. The variation of the spectrum of a normal matrix. — Duke Math. J., 1953, 20, p. 37–39.
101. Householder A.S. Unitary triangularization of a nonsymmetric matrix. — J. ACM, 1958, 5, p. 339–342.
102. Householder A.S. The theory of matrices in numerical analysis. — New York: Blaisdell Publ. Co., 1964.
103. Householder A.S. KWIC index for numerical algebra. — Oak Ridge Nat. Lab. Rept. № ORNL-4778, 1972, 538 p.
104. Jacobi C.G.J. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. — Crelle's J., 1846, 30, p. 297–306.
105. Jennings L.S., Osborn M.R. Applications of orthogonal matrix transformations to the solution of systems of linear and nonlinear equations. — Australian Nat. Univ. Comput. Centre Tech. Rept. № 37. — Australia, Canberra, 1970, 45 p.
106. Jennrich R.I., Sampson P.I. Remark AS-R3. A remark on algorithm AS-10. Appl. Stat., 1971, 20, p. 117–118.
107. Kahan W. When to neglect off-diagonal elements of symmetric tri-diagonal matrices. — Stanford Univ. Rept. № CS-42. — Calif., Stanford, 1966.

108. Kahan W. Numerical linear algebra. — Canad. Math. Bull., 1966, 9, № 6, p. 757–801.
109. Kalman R.E. A new approach to linear filtering and prediction problems. — ASME Trans., J. Basic Eng., 1960, 82D, p. 35–45.
110. Kammerer W.J., Nashed M.Z. On the convergence of the conjugate gradient method for singular linear operation equations. — SIAM J. Numer. Anal., 1972, 9, p. 165–181.
111. Korganoff A., Pavel-Parvu M. Éléments de théorie des matrices carrées et rectangles an analyse numérique. — Paris: Dunod, 1967.
112. Krogg F.T. Efficient implementation of a variable projection algorithm for nonlinear least squares problems. — Commun. ACM, 1974, 17, № 3, p. 167–169.
113. La Motte L.R., Hocking R.R. Computational efficiency in the selection of regression variables. — Technometrics, 1970, 12, p. 83–93.
114. Lawson C.L. Contributions to the theory of linear least maximum approximation. — Thesis. UCLA. — Calif., Los Angeles, 1961, 99 p.
115. Lawson C.L. Applications of singular value analysis. — Mathematical Software. — New York: Academic Press, 1971, p. 347–356.
116. Leringe Ö., Wedin P.-Å. A comparison between different methods to compute a vector  $x$  which minimizes  $\|Ax - B\|$ , when  $Gx = h$ . — Lund. Univ. — Sweden, Lund, 1970, 21 p.
117. Levenberg K. A method for the solution of certain nonlinear problems in least squares. — Quart. Appl. Math., 1944, 2, p. 164–168.
118. Longley J.W. An appraisal of least squares programs for the electronic computer from the point of view of the user. — J. Amer. Stat. Ass., 1967, 62, p. 819–841.
119. Lynn M.S., Timlake W.P. The use of multiple deflations in the numerical solution of singular systems of equations with applications to potential theory. — IBM Houston Sci. Center, 37.017. — Texas, Houston, 1967.
120. Marcus M., Gordon W.R. An analysis of equality in certain matrix inequalities II. — SIAM J. Numer. Anal., 1972, 9, p. 130–136.
121. Marquardt D.W. An algorithm for least-squares estimation of nonlinear parameters. — J. Soc. Indust. Appl. Math., 1963, 11, № 2, p. 431–441.
122. Marquardt D.W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. — Technometrics, 1970, 12, p. 591–612.
123. Mason J.C. Orthogonal polynomial approximation methods in numerical analysis. — Univ. of Toronto Tech. Rept. № 11. — Canada, Ontario, Toronto, 1969, 50 p.
124. Meyer C.D., Jr. The Moore-Penrose inverse of a bordered matrix. — Lin Alg. and Its Appl., 1972, 5, p. 375–382.
125. Mitchell W.C., McCraith D.L. Heuristic analysis of numerical variants of the Cram-Schmidt orthonormalization process. — Stanford Univ. Rept. № CS-122. — Calif., Stanford, 1969.
126. Moler C.B. Iterative refinement in floating point. — J. ACM, 1967, 14, № 2, p. 316–321.
127. Moler C.B., Stewart G.W. An algorithm for generalized matrix eigenvalue problem. — SIAM J. Numer. Anal., 1973, 10, № 2, p. 241–256.
128. Moore E.H. On the reciprocal of the general algebraic matrix. — Bulletin AMS, 1920, 26, p. 394–395.
129. Moore E.H. General analysis, Part 1. — Mem. Amer. Philos. Soc., 1935, 1, p. 1–231.
130. Morrison D.D. Methods for nonlinear least squares problems and convergence proofs. — Proc. of Seminar on Tracking Programs and Orbit Determination. — Calif., Pasadena, Jet Propulsion Lab., 1960, p. 1–9.
131. Murray W. An algorithm to find a local minimum of an indefinite quadratic program. — Nat. Phys. Lab. Rept. № NPL-DNAC-1. — England, Middlesex, Teddington, 1971, 31 p.
132. Newman M., Todd J. The evaluation of matrix inversion programs. — J. Soc. Indust. Appl. Math., 1958, 6, № 4, p. 466–476.
133. Osborne E.E. On least squares solutions of linear equations. — J. ACM, 1961, 8, p. 628–636.
134. Osborne E.E. Smallest least squares solutions of linear equations. — SIAM J. Numer. Anal., 1965, 2, № 2, p. 300–307.

135. Paige C.C. An error analysis of a method for solving matrix equations. — Stanford Univ. Rept. № CS-297. — Calif., Stanford, 1972.
136. Parlett B.N. The LU and QR transformations. — Mathematical methods for digital computers, II. — New York: Wiley, 1967, p. 116–130.
137. Pavel-Parvu M., Korganoff A. Iteration functions for solving polynomial matrix equations. — Constructive aspects of the fundamental theorem of algebra. — New York: Wiley, 1969, p. 225–280.
138. Penrose R. A generalized inverse for matrices. — Proc. Cambridge Phil. Soc., 1955, 51, p. 406–413.
139. Pereyra V. Stabilizing linear least squares problems. — Proc. IFIP Congr. 1968. — Amsterdam: North-Holland, 1969, p. 119–121.
140. Pereyra V. Stability of general systems of linear equations. — Aequat. Math., 1969, 2, № 2–3, p. 194–206.
141. Perez A., Scolnik H.D. Derivatives of pseudoinverses and constrained nonlinear regression problems. — Numer. Math.
142. Peters G., Wilkinson J.H. The least squares problem and pseudoinverses. — Comput. J., 1970, 13, p. 309–316.
143. Philips D.L. A technique for the numerical solution of certain integral equations of the first kind. — J. ACM, 9, p. 84–97.
144. Plackett R.L. Principles of regression analysis. — New York: Oxford Univ. Press, 1960.
145. Powell M.J.D. A Fortran subroutine for solving systems of nonlinear algebraic equations. — Atomic Energy Res. Estab. Rept. № R-5947. — England, Berkshire, Harwell, 1968.
146. Powell M.J.D., Reid J.K. On applying Householder's method to linear least squares problems. — Atomic Energy Res. Estab. Rept. № T-P-332. — England, Berkshire, Harwell, 1968, 20 p.
147. Powell M.J.D., Reid J.K. On applying Householder's method to linear least squares problems. — Proc. IFIP Congr. 1968. — Amsterdam: North-Holland, 1969, p. 122–126.
148. Powell M.J.D. A survey of numerical methods for unconstrained optimization. — SIAM Rev., 1970, 12, p. 79–97.
149. Pringle R.M., Rayner A.A. Expressions for generalized inverses of a bordered matrix with application to the theory of constrained linear models. — SIAM Rev., 1970, 12, p. 107–115.
150. Pyle L.D. A generalized inverse (epsilon)-algorithm for constructing intersection projection matrices, with applications. — Numer. Math., 1967, 10, p. 86–102.
151. Ralston A., Wilf H.S. Mathematical methods for digital computers. — New York: Wiley, 1960.
152. Ralston A., Wilf H.S. Mathematical methods for digital computers, II. — New York: Wiley, 1967.
153. Rao C.R., Mitra S.K. Generalized inverse of matrices and its application. — New York: Wiley, 1971.
154. Reid J.K. A Fortran subroutine for the solution of large sets of linear equations by conjugate gradients. — Atomic Energy Res. Estab. Rept. № 6545. — England, Berkshire, Harwell, 1970, 5 p.
155. Reid J.K. The use of conjugate gradients for systems of linear equations possessing property A. — SIAM J. Numer. Anal., 1972, 9, № 2, p. 325–332.
156. Reid J.K. (ed.). Large sparse sets of linear equations. — New York: Academic Press, 1971.
157. Rice J.R. Experiments on Gram-Schmidt orthogonalization. — Math. Comp., 1966, 20, p. 325–328.
158. Rice J.R. The approximation of functions, 2 — advanced topics. — Addison Wesley Publ. Co., 1969.
159. Rice J.R. Running orthogonalization. — J. Approximation Theory, 1971, 4, p. 332–338.
160. Rice J.R. (ed.). Mathematical software. — New York: Academic Press, 1971.
161. Rice J.R., White J.S. Norms for smoothing and estimation. — SIAM Rev., 1964, 6, № 3, p. 243–256.
162. Rosen E.M. The instrument spreading correction in GPC III. The general shape function using singular value decomposition with a nonlinear calibration curve. — Monsanto Co., St. Louis, Mo., 1970.
163. Rosen J.B. The gradient projection method for nonlinear programming, Part I, linear constraints. — J. Soc. Indust. Appl. Math., 1960, 8, № 1, p. 181–217.
164. Rutishauser H. Once again the least squares problem. — Lin. Alg. and Its Appl., 1968, 1, p. 479–488.
165. Saunders M.A. Large-scale linear programming using the Cholesky factorization. — Stanford Univ. Rept. CS-252. — Calif., Stanford, 1972, 64 p.
166. Saunders M.A. Product form of the Cholesky factorization for large-scale linear programming. — Stanford Univ. Rept. CS-301. — Calif., Stanford, 1972, 38 p.
167. Schur I. Über die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen. — Math. Ann., 1909, 66, p. 488–510.
168. Sherman J., Morrison W.J. Adjustment of an inverse matrix corresponding to the changes in the elements of a given column or a given row of the original matrix. — Ann. Math. Stat., 1949, 20, p. 621.
169. Sherman J., Morrison W.J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. — Ann. Math. Stat., 1950, 21, № 1, p. 124–127.
170. Smith G.L. On the theory and methods of statistical inference. — NASA Tech. Rept. TR R-251. — Washington, 1967, 32 p.
171. Stewart G.W. On the continuity of the generalized inverse. — SIAM J. Appl. Math., 1969, 17, № 1, p. 33–45.
172. Stewart G.W. Incorporating origin shifts into the QR algorithm for symmetric tridiagonal matrices. — Commun. ACM, 1970, 13, № 6, p. 365–367.
173. Stewart G.W. Introduction to matrix computations. — New York: Academic Press, 1973.
174. Stoer J. On the numerical solution of constrained least squares problems. — SIAM J. Numer. Anal., 1971, 8, № 2, p. 382–411.
175. Strand O.N., Westwater E.R. Statistical estimation of the numerical solution of a Fredholm integral equation of the first kind. — J. ACM, 1968, 15, № 1, p. 100–114.
176. Strand O.N., Westwater E.R. Minimum-RMS estimation of the numerical solution of a Fredholm integral equation of the first kind. — SIAM J. Numer. Anal., 1968, 5, № 2, p. 287–295.
177. Swerling P. A proposed stagewise differential correction procedure for satellite tracking and prediction. — J. Astronaut. Sci., 1959, 6, № 3, p. 46–52.
178. Tornheim L. Stepwise procedures using both directions. — Proc. 16th Nat. Meeting of ACM, 12A4.1–12A4.4.
179. Tucker A.W. Least distance programming. — Proc. of the Princeton Sympos. on Math. Programming. — Princeton: Princeton Univ. Press, p. 583–588.
180. Turing A.M. Rounding-off errors in matrix processes. Quart. J. Mech., 1948, 1, p. 287–308.
181. Twomey S. On the numerical solution of Fredholm integral equations of the first kind by inversion of the linear system produced by quadrature. — J. ACM, 1963, 10, p. 97–101.
182. Van der Sluis. Condition numbers and equilibrium of matrices. — Numer. Math., 1969, 14, p. 14–23.
183. Van der Sluis. Stability of solutions of linear algebraic systems. — Numer. Math., 1970, 14, p. 246–251.
184. Varah J.M. Computing invariant subspaces of a general matrix when the eigensystem is poorly conditioned. Univ. of Wisconsin Math. Res. Center Rept. 962. — Wis., Madison, 1969, 22 p.
185. Wampler R.H. An evaluation of linear least squares computer programs. — Nat. Bur. of Standards J. Res., 1969, 73B, № 2, p. 59–90.
186. Wedin P.-A. Perturbation bounds in connection with singular value decomposition. — BIT, 1972, 12, p. 99–111.
187. Wedin P.-A. Perturbation theory for pseudo-inverses. — BIT, 1973, 13, p. 217–232.
188. Wedin P.-A. On the almost rank deficient case of the least squares problem. — BIT, 1973, 13, p. 344–354.

189. Westwater E.R., Strand O.N. Statistical information content of radiation measurements used in indirect sensing. — J. Atmospheric Sciences, 1968, 25, № 5, p. 750–758
190. Wilkinson J.H. Error analysis of floating-point computation. — Numer. Math., 1960, 2, p. 319–340.
191. Wilkinson J.H. Householder's method for symmetric matrices. — Numer. Math., 1962, 4, p. 354–361.
192. Wilkinson J.H. Rounding errors in algebraic processes. — Prentice-Hall, 1963.
193. Wilkinson J.H. Convergence of the LR, QR and related algorithms. — Comput. J., 1965, 8, № 1, p. 77–84.
194. Wilkinson J.H. Global convergence of tridiagonal QR algorithm with origin shifts. Lin. Alg. and Its Appl., 1968, 1, p. 409–420.
195. Wilkinson J.H. Global convergence of QR algorithm. — Proc. IFIP Congr. 1968. — Amsterdam: North-Holland, 1969, p. 130–133.
196. Wilkinson J.H. Elementary proof of the Wielandt-Hoffman theorem and its generalization. Stanford Univ. Rept. NCS-150. — Calif., Stanford, 1970.
197. Willoughby R.A. (ed.). Sparse matrix proceedings. — Thomas J. Watson Res. Center Rept. RAI-(11707), 1968, 184 p.
198. Wolfe P. The composite simplex algorithm. — SIAM Rev., 1965, 7, p. 42–54.

## СПИСОК ЛИТЕРАТУРЫ, ДОБАВЛЕННОЙ ПРИ ПЕРЕВОДЕ

- 1\*. Воеводин В.В. Ортогональные преобразования и решение систем уравнений с прямоугольными матрицами. — В кн.: Ошибки округления в алгебраических процессах. — М.: Изд-во МГУ, 1968, с. 39–58.
- 2\*. Дрыгалла Ф. Уточнение псевдорешений системы линейных уравнений. — Журнал вычисл. мат. и матем. физ., 1982, 22, № 5, с. 1–7.
- 3\*. Икрамов Х.Д. Численное решение линейных задач метода наименьших квадратов. — В кн.: Математический анализ, 23. Итоги науки и техники, М.: ВИНТИ, 1985.
- 4\*. Кублановская В.Н. О вычислении обобщенной обратной матрицы и проектора. — Журнал вычисл. мат. и матем. физ., 1966, 6, № 2, с. 326–332.
- 5\*. Кублановская В.Н. О применении метода Ньютона к определению собственных значений матриц. — Докл. АН СССР, 1969, 188, № 5, с. 1004–1005.
- 6\*. Кублановская В.Н. Применение ортогональных преобразований к численной реализации линейно алгебраических задач на возмущение. — Журнал вычисл. мат. и матем. физ., 1970, 10, № 2, с. 429–433.
- 7\*. Кублановская В.Н. Применение нормализованного процесса к решению обратной проблемы собственных значений матриц. — Зап. науч. семинаров ЛОМИ АН СССР, 1971, 23, с. 72–83.
- 8\*. Кублановская В.Н. Применение нормализованного процесса к решению линейных алгебраических систем. — Журнал вычисл. мат. и матем. физ., 1972, 12, № 5, с. 1091–1098.
- 9\*. Парлетт Б. Симметричная проблема собственных значений. Численные методы. — М.: Мир, 1983.
- 10\*. Уилкинсон Дж., Райнш К. Справочник алгоритмов на языке АЛГОЛ. Линейная алгебра. — М.: Машиностроение, 1976.
- 11\*. Фаддеев Д.К., Кублановская В.Н., Фаддеева В.Н. Линейные алгебраические системы с прямоугольными матрицами. — В кн.: Современные численные методы. Вып. I (Материалы междунар. летней школы по численным методам. Киев, 1966). — М., 1968, с. 16–75.
- 12\*. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. — М.: Мир, 1980.
13. Bauer F.L. Elimination with weighted row combinations for solving linear equations and least squares problems. — Numer. Math., 1965, 7, p. 338–352.

14. Björck A. Iterative refinement of linear least squares solutions, I, II. — BIT, 1967, 7, p. 257–278; 1968, 8, p. 8–30.
15. Björck A. Comment on the iterative refinement of least squares solutions. — J. Amer. Statist. Assoc., 1978, 73, p. 161–166.
16. Björck A., Duff I.S. A direct method for the solution of sparse linear least squares problems. — Linear Algebra and Appl., 1980, 34, p. 43–67.
17. Björck A., Golub G.H. Iterative refinement of linear least squares solutions by Householder transformation. — BIT, 1967, 7, p. 322–337.
18. Businger P., Golub G.H. Linear least squares solutions by Householder transformations. — Numer. Math., 1965, 7, p. 269–276.
19. Chan T.F. An improved algorithm for computing the singular value decomposition. — ACM Trans. Math. Software, 1982, 8, № 1, p. 72–88.
20. Cline A.K. An elimination method for the solution of linear least squares problems. — SIAM J. Numer. Anal., 1973, 10, № 2, p. 283–289.
21. Cline A.K., Conn A.R., Van Loan Ch. F. Generalizing the LINPACK condition estimator. — Lect. Notes Math., 1982, № 909, p. 73–83.
22. Cline A.K., Moler C.B., Stewart G.W., Wilkinson J.H. An estimate for the condition number of a matrix. — SIAM J. Numer. Anal., 1979, 16, № 2, p. 368–375.
23. Cline R.E., Plemmons R.J.  $l_1$ -solutions to undetermined linear systems. — SIAM Review, 1976, 18, p. 92–106.
24. Cuppen J.J.M. The singular value decomposition in product form. — SIAM J. Sci. Stat. Comp., 1983, 4, № 2, p. 216–222.
25. Davis Ch.D., Kahan W.M. The rotation of eigenvectors by a perturbation. — SIAM J. Numer. Anal., 1970, 7, p. 1–46.
26. Deuffhard P., Sautter W. On rank-deficient pseudoinverses. — Linear Algebra and Appl., 1980, 29, p. 91–111.
27. Elden L. Perturbation theory for the least squares problem with linear equality constraints. — SIAM J. Numer. Anal., 1980, 17, p. 338–350.
28. Elden L. A weighted pseudoinverse, generalized singular values and constrained least squares problems. — BIT, 1982, 22, p. 487–502.
29. Fletcher R.H. On the iterative refinement of least squares solutions. — J. Amer. Statist. Assoc., 1975, 70, p. 109–112.
30. Forsythe G.E., Golub G.H. On the stationary values of a second degree polynomial on the unit sphere. — J. Soc. Indust. Appl. Math., 1965, 13, p. 1050–1068.
31. Gander W. Least squares with a quadratic constraint. — Numer. Math., 1981, 36, p. 291–307.
32. Golub G.H. Numerical methods for solving linear least squares problems. — Numer. Math., 1965, 7, № 3, p. 206–216.
33. Golub G.H. Matrix decompositions and statistical calculations. — In: Statistical Computation. — New York: Academic Press, 1969, p. 365–397.
34. Golub G.H., Van Loan Ch.F. Matrix computations. — The Johns Hopkins University Press, 1983.
35. Golub G.H., Wilkinson J.H. Note on the iterative refinement of least squares solution. — Numer. Math., 1966, 9, p. 139–148.
36. Hammarling S.J., Long E.M.R., Martin D.W. A generalized linear least squares algorithm for correlated observations, with special reference to degenerate data. — Nat. Phys. Lab. Inf. Technol. and Comput. Rept., 1983, № 33, 88 p.
37. Hestenes M.R. Inversion of matrices by biorthogonalization and related results. — J. Soc. Indust. Appl. Math., 1958, 6, p. 51–90.
38. Jennings L.S., Osborne M.R. A direct error analysis for least squares. — Numer. Math., 1974, 22, p. 325–332.
39. Jiguang, Sun. Perturbation analysis for the generalized eigenvalue and the generalized singular value problem. — In: Matrix Pencils. Springer-Verlag, 1983, p. 221–244.
40. Kaneko I., Plemmons R.J. Minimum norm solutions to linear elastic analysis problems. — Int. J. Numer. Meth. Eng., 1984, 20, № 6, p. 983–998.
41. Karasalo I. A criterion for truncation of the QR-decomposition algorithm for the singular linear least squares problem. — BIT, 1974, 14, p. 156–166.
42. Kourouklis S., Paige C.C. A constrained squares approach to the general Gauss-Markov linear model. — J. Amer. Stat. Assoc., 1981, 76, p. 620–625.

43. Leringe Ö., Wedin P.A. A comparison between different methods to compute a vector  $x$ , which minimizes  $\|Ax-B\|$ , when  $Gx=h$ . — Sweden, Lund, Lund University, 1970, 21 p.
44. Lötstedt P. Perturbation bounds for the linear least squares problem subject to linear inequality constraints. — BIT, 1983, 23, № 4, p. 500–519.
45. Lötstedt P. Solving the minimal least squares problem subject to bounds on the variables. — BIT, 1984, 24, № 2, p. 206–224.
46. Luk F.T. Computing the singular-value decomposition on the ILLIAC IV. — ACM Trans. Math. Software, 1980, 6, № 4, p. 524–539.
47. Mitra S.K., Rao C.R. Projections under seminorms and generalized Moore-Penrose inverses. — Linear Algebra and Appl., 1974, 9, p. 155–167.
48. Nash J.C. A one-sided transformation method for the singular value decomposition and algebraic eigenproblem. — Comput. J., 1975, 18, p. 74–76.
49. Nashed M.Z. Generalized inverses and applications. — New York: Academic Press, 1976.
50. Paige C.C. An error analysis of a method for solving matrix equations. — Math. Comp., 1973, 27, № 122, p. 355–359.
51. Paige C.C. Numerically stable computations for general univariate linear models. — Comm. Statist., 1978, B7, № 5, p. 437–453.
52. Paige C.C. Computer solution and perturbation analysis of generalized least squares problems. — Math. Comp., 1979, 33, p. 171–183.
53. Paige C.C. Fast numerically stable computations for generalized linear least squares problems. — SIAM J. Numer. Anal., 1979, 16, p. 165–171.
54. Paige C.C. Error analysis of some techniques for updating orthogonal decompositions. — Math. Comput., 1980, 34, № 150, p. 465–471.
55. Paige C.C., Saunders M.A. Towards a generalized singular value decomposition. — SIAM J. Numer. Anal., 1981, 18, p. 398–405.
56. Plemmons R.J. Linear least squares by elimination and MGS. — J. ACM, 1974, 21, № 4, p. 581–585.
57. Sautter W. Zur Kondition des linearen Ausgleichsproblems mit linearen Gleichungen als Nebenbedingungen. — Numer. Math., 1984, 44, p. 139–152.
58. Schittkowski K. The numerical solution of constrained linear least-squares problems. — IMA J. Numer. Anal., 1983, 3, № 1, p. 11–36.
59. Slius A. Stability of the solutions of linear least squares problems. — Numer. Math., 1974/75, 23, p. 241–254.
60. Stewart G.W. A note on the perturbation of singular values. — Linear Algebra and Appl., 1979, 28, p. 213–216.
61. Stewart G.W. A method for computing the generalized singular value decomposition. — In: Matrix Pencils. Springer-Verlag, 1983, p. 207–220.
62. Stewart G.W. A second order perturbation expansion for small singular values. — Linear Algebra and Appl., 1984, 56, p. 231–235.
63. Tsao N.K. On the orthogonal factorization and its updating in band-structured matrix computations. — J. Franklin Inst., 1981, 311, № 6, p. 355–381.
64. Van Loan Ch.F. Generalizing the singular value decomposition. — SIAM J. Numer. Anal., 1976, 13, p. 76–83.
65. Van Loan Ch.F. A generalized SVD analysis of some weighting methods for equality constrained least squares. — In: Matrix Pencils. Springer-Verlag, 1983, p. 245–262.
66. Wilkinson J.H. Rounding errors in algebraic processes. — Prentice-Hall, 1963.

## ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

### Алгоритм Пауэлла-Рида 80, 115

- BSEQT 165
- COV 54
- G1 46
- G2 46
- H1 45
- H2 45
- HBT 58
- HFT 50
- HFTI 62
- HSI 50
- HS2 58
- LDP 127
- LSE 107
- NNLS 124
- QRBD 88
- SEQHT 163

### Арифметика со смешанной точностью 76

### Взвешивание 114

### Возмущения класса I 198

### — II 199

### Вытеснение 87

### Гиперплоскость 183

### Главный элемент преобразования Хаусхолдера 43

### Гребневая регрессия 145

### Двойственный вектор задачи НКН 124

### Задача наименьших квадратов (задача НК) 8

### — с квадратичным ограничением (задача НККО) 216

### — с ограничениями-неравенствами (задача НКН) 122

### — с ограничениями-уравнениями (задача НКУ) 58, 103

### — о выборе подмножества 149

### — LDP 122

### — NNLS 122

### Итерационное уточнение 201

### Линейное многообразие 182

### Матрица вращения 184

### — двухдиагональная 183

### — диагональная 183

### — единичная 183

### — идемпотентная 185

### — Кахана 25, 27

### — ковариационная 52

### — обратная 183

### — левая 183

### — правая 183

### — ортогональная 183

### — отражения 184

### — перестановки 184

### — проекционная 185

### — псевдообратная 31

### — ML-взвешенная 19898

### — симметричная 184

### — неотрицательно определенная 184

### — положительно определенная 184

### — транспонированная 180

### — треугольная 183

### — Хаусхолдера 12

### — хессенбергова 183

### — эквивалентного возмущения 212

### Матрицы главная диагональ 183

### — диагональные элементы 183

### — инвариантное подпространство 184

### — симметричной собственное подпространство 184

### — собственные векторы 184

### — значения 184

### — спектральное разложение 184

### Метод Джентльмена 47

### — Клайна 206, 207

### — Марквардта 145

### — односторонней ортогонализации 205

### — Питерса-Уилкинсона 102, 206

### — стабилизации Левенберга-Марквардта 152

### — Холесского 94, 95

### — без квадратных корней 103

### Модифицированная ортогонализация

### Грама-Шмидта 92



Модифицированный метод Грама—Шмид-  
та 100

Норма вектора евклидова 180  
— матрицы евклидова (Шура, Фро-  
бениуса) 181

— спектральная 181

Нормализованный процесс 200

Нормальное псевдорешение (решение за-  
дачи НК) 39

— решение (решение минимальной дли-  
ны) 39

Нормальные уравнения 92, 216

Оболочка системы векторов 182

Образ матрицы 182

Ортогональное разложение матрицы 11

Подпространства базис 181

— ортогональное дополнение 182

— размерность 181

Подпространство 181

Полупространство 183

Последовательное накопление 161

— оценивание 160

Пошаговая регрессия 150

Преобразование Гивенса 12

— Хаусхолдера 12

Пробное решение 151

Пространство столбцов матрицы 182

— строк матрицы 182

Прямая сумма подпространств 182

Псевдоранг 59

Размерность линейного многообразия 182

— подпространства 181

Ранг матрицы 182

— — неполный 182

— — полный 182

Сингулярное разложение матрицы 18

— — пары матриц 194

Сингулярные числа 18

Сингулярный анализ 151

Система векторов линейно зависима 181

— — — независима 181

— — ортогональная 183

Сплайн кубический 172

— линейный 169

Теорема Виландта—Хофмана 22

— о глобальной сходимости  $QR$ -  
ритма 83

Теоремы Пауэлла—Рида 79—81

Условия Куна—Таккера 123

— Пенроуза 32

Фильтрация 160

Число обусловленности 39

Ширина ленты 164

Ядро матрицы 182

$CS$ -разложение ортонормальной матри-  
цы 195

$QR$ -алгоритм 82

$QR$ -разложение матрицы 12

$QR$ -разложения модификация 134, 160,  
174

Чарльз Лоусон, Ричард Хенсон

## ЧИСЛЕННОЕ РЕШЕНИЕ ЗАДАЧ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Редактор *Е.Е. Тыртышников*

Художественный редактор *Т.Н. Колесникова*

Технический редактор *С.В. Геворкян*

Корректор *Т.В. Обод*

Набор осуществлен в издательстве  
на наборно-печатающих автоматах

ИБ № 12433

Сдано в набор 09.12.85 Подписано к печати 14.03.86

Формат 60 X 90 1/16. Бумага офсетная

Гарнитура Пресс-Роман. Печать офсетная. Усл.печ.л. 14,5

Усл.кр.-отт. 14,5. Уч.-издл. 15,2. Тираж 12 300 экз.

Тип. зак. 25. Цена 1 р. 40 к.

Ордена Трудового Красного Знамени  
издательство "Наука"

Главная редакция физико-математической литературы  
117071 Москва В-71, Ленинский проспект, 15

4-я типография издательства "Наука"

630077 г. Новосибирск-77, ул. Станиславского, 25